Commentary

MANPOWER AND CULTURE

MANPOWER used to be measured by counting heads. Nowadays it is important what is inside the heads. Napoleon is supposed to have said, 'Only numbers annihilate', and to have invented the mass levy though it is doubtful if the concept or the practice would have been regarded as novel by Genghiz Khan. In the last great war we were told that it required eleven men behind the fighting soldier to keep him fighting. In times of war there is an unanimity of immediate purpose which we rarely achieve in peacetime, partly because the *immediate* purpose is paramount. The parameters thus simply and starkly laid bare in war, remain coercive in peace. It is a dismal way to assess human affairs in terms of conflict; we are effete if we pretend these factors do not exist.

A culture is a cultivating. It is more than the climate, it is the whole organic process, spiritual and material, aesthetic and functional. In the last two hundred years we have developed two cultures which appear to some degree to have become polarized, the humanist and the scientific, and instead of becoming integrated, the two are becoming more polarized. Sir Herbert Read, as President of the Society for Education through Art is reported to have said[1]:

> 'The ideal of technology is complete automation; the corresponding ideal of education is a human brain that controls itself, free from emotional or idealistic entanglements, free, above all, from originality of any kind. Functional thinking is like functional machinery, cold, precise, imageless, repetitive, bloodless, nerveless, dead ... For a century or two we may be able to control its devastating effects by establishing a universal tyranny more absolute than any yet experienced by mankind. But disaster, as the scientists themselves have predicted, is now a mathematical certainty.'

Such despair is so wildly ludicrous to anyone with any appreciation of the scientific culture that it is brushed aside as incomprehensible.

This despair, this negative preciosity should not be brushed aside. It is a widely held view and it is thoroughly alarming for the circumstances of which Read complains are the very circumstances of *his* opportunity. People who are terrified of survival have little time for the enriching of life. The chance, the opportunity, is the leisure that power over natural resources brings. The population of the British Isles was 19 million in 1815 and of the years that followed Ramsay Muir writes[2]:

> 'It is the most poignant commentary on the condition of Britain that, according to the state of trade, from one tenth to one fourth of the population of England and Wales were paupers, drawing allowances from the poor-rates to supplement their wages because these were insufficient to maintain life. What was worst of all, by the receipt of poor-relief, men were forced to submit themselves and their families to a sort of slavery to the poor law authorities . . . British liberty had become a very unreal thing for those who were subjected to such conditions.'

We have many many problems to occupy our attention today but there have been some changes. Dr MARK ABRAMS has recently prepared a survey of teenagers' spending[3]; the average British youth has £5 a week to spend and his girl £3. They form a very important market.

The purpose of education is to widen horizons and an educational system is a reflection and an expression of a culture.

> 'The curious thing was that in Germany, in the 1830s and 1840s, long before serious industrialization had started there, it was possible to get a good university education in applied science, better than anything England or the United States could offer for a couple of generations. I do not begin to understand this; it does not make social sense: but it was so, with the result that Ludwig Mond, the son of a court purveyor, went to Heidelberg and learnt some sound applied chemistry. Siemens, a Russian signals officer, at military academy and university went through what for their time were excellent courses in electrical engineering. Then they came to England, met no competition at all, brought in other educated Germans and made fortunes exactly as though they were dealing with a rich, illiterate colonial territory[4].'

Sir Charles Snow has called his Rede Lecture[4] 'The Two Cultures and the Scientific Revolution'. Certainly the two have much to learn from each other. At present through ignorant jealousy each copies the worst features of the other. There is a pride in incomprehension which is simply disgraceful—and stems from laziness and fear. Soon a competent journeyman will have the leisure of a landed aristocrat and will need the equivalent training of a University General Degree. Without the technical training he will be illiterate; if his leisure is not developed, he will be 'as the beasts that perish'. Our present educational system is a patchwork of legacies and pressures. The design of a patchwork has the real merit of flexibility; there is much that is good but there is a lot of lumber. What is needed is a radical re-thinking of the purposes, and this is really the problem for there is no single purpose. The purposes are as diverse as the culture it should reflect.

[1] See *Daily Telegraph* (23rd April, 1957);

[2] MUIR, R. *Short History of the British Commonwealth*, p. 314 London: Phillip, 1934

[3] *Manchester Guardian* (1st August, 1959);

[4] SNOW, C. P. *The Two Cultures and The Scientific Revolution* London: Cambridge University Press, 1959

THE USE OF CERAMIC MATERIALS FOR GAS TURBINE BLADES

J. M. PURI

Imperial Chemical Industries (India) Ltd, Calcutta

The article is a survey of the use of ceramic materials for gas turbine blades and is based partly on the author's own work at Imperial College, London, and partly on a paper presented by T. H. BLAKELEY and R. F. DARLING to the North East Coast Institution of Engineers and Shipbuilders. This paper was one of the entries in the 1958 Waverley Gold Medal Essay Competition.

THE outstanding need to raise the operating temperatures, and thereby the thermal efficiency, of the gas turbine has given tremendous impetus to the search for satisfactory materials of construction for fabricating blades and other component parts. For a number of years work has been done on the development of gas turbines employing inlet temperatures of 1200°C or more. These temperatures are well beyond what could be withstood by metallic materials, and hence attention has to be turned to non-metals, *i.e.*, ceramics. Furthermore the possibility that certain types of ceramics might meet this need has focussed attention on the general question of ceramics as engineering materials.

At temperatures of the order of 1000°C, metals often fail as materials of construction through low strength, low creep resistance or poor resistance to oxidation and, more generally, their pronounced instability to other various forms of atmospheric attack. In comparison with metals, ceramics are immediately attractive, because of their freedom from oxidation and their chemical stability at high temperature. Their low density (seldom exceeding four) is also an attractive feature to minimize centrifugal stresses in parts rotating at high speed.

Properties and Limitations of Ceramics

Besides the above mentioned advantages of ceramics various other properties may be considered. All common commercial refractories and ceramics are based on mixtures of some major and several minor components. During the burning process, needed to produce the ceramic, eutectics are formed which melt to give liquid in addition to the solid crystalline phases and when the material is cooled, the liquid often persists as glass. The strength of this type of ceramic, at both normal and high temperatures, is due to the inherent strength of both crystals and the glass or liquid bond. As the temperature is raised, the properties of the liquid phase soon become dominant, and as a result the material changes from a state of perfect elasticity to one of plasticity. For most commercial materials this transition is effective below 1000°C, when the material begins to show pronounced creep, which is centered in the liquid phase, but probably much modified by the crystalline components.

Whilst the strength of many materials of this type is adequate at room temperature, it is soon lost on heating and, in addition, creep becomes excessive. The corollary is that ceramics of suitable strength and creep resistance at high temperature will be found only in two types of material:

(*i*) multi-component materials forming eutectics of very high melting point, and

(*ii*) single-component high melting point materials giving theoretically no liquid below the melting point of the component itself.

Class (*i*) consists mainly of the less common materials with melting points in the range 2000° to 3000°C, and much more information on the properties of such mixtures is required. On the other hand, there is now a good deal of knowledge and experience of class (*ii*) materials, which are simpler to obtain and to study.

Before specific materials are described, further generalizations about ceramics should be mentioned. Their main drawbacks as engineering materials include brittleness, that is the virtual absence of ductility, and abrasiveness. In general they cannot be machined or built up from stock, although the shaping or turning of unfired or lightly fired material is sometimes possible with small articles. Usually the materials must be fabricated by what are, in effect, powder methods and then burned at high temperatures[1].

The low ductility is a serious disadvantage in several respects. Such materials are liable to break

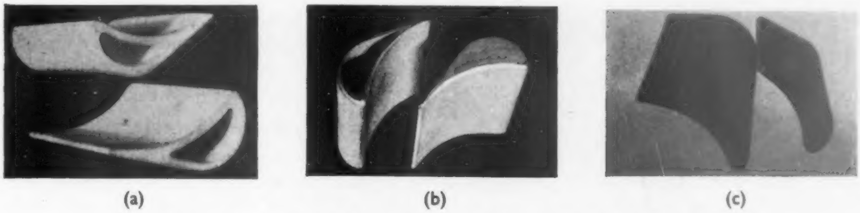


Figure 1. (LEFT) *Alumina blades;* (CENTRE) *hollow and solid blade made from zircon–chromic oxide; and* (RIGHT) *blades made from IC silicon carbide*

down as the result of local stress concentrations—the ability of metals to yield and distribute high localized stresses being notoriously absent. Whilst the tensile strength of certain sintered oxides appears to be high enough to meet estimated stresses under turbine conditions, these high values are obtained only when great care is taken in testing to ensure axial loading and thus to avoid local stress concentration[2]. In other words, failure in practice of intricate shapes such as turbine blades might often occur even though laboratory tests indicate that the strength of the material is adequate. Similarly, the low impact strength is an added danger, as materials that are liable to fly to pieces, rather than to bend, on collision with particles in the gas stream, clearly could not be tolerated.

Materials Considered

Numerous ceramic materials have adequate hot strength but poor thermal shock resistance. The fundamental physical properties that favour good thermal shock resistance are: low coefficient of expansion, low Young's modulus and high strength, and high thermal conductivity. These criteria provide an initial sorting test for possible materials; they ignore various textural factors which are less easy to assess quantitatively.

The materials considered of interest were beryllia, alumina, zircon, silicon carbide, and metal–ceramic bodies (sometimes known as 'cermets'). Of these, beryllia was not considered in this investigation[3] on account of its toxicity and attendant complications in manufacture. Before going on to describe the actual tests performed, a brief account will be given of the materials.

Alumina—Recrystallized alumina was used as a blade material. Its thermal shock resistance was considerably improved by incorporating relatively coarse material; this technique broke up the continuous structure. *Figure 1* (LEFT) shows two blades, the bigger of the two blades was coarse alumina (bonded with silica), and the other blade recrystallized alumina.

Zircon—The properties of zircon (zirconium silicate) indicate that it is a promising material. Preliminary tests on zircon showed that it was relatively weak at 700° to 800°C, as compared with higher or lower temperatures. It was found that this defect could be eliminated by small additions of chromic oxide Cr_2O_3 with an optimum at about two and a half per cent addition (see *Figure 1*).

An unfortunate property of zircon is its tendency to dissociation into zirconia and silica at, or near, its firing temperature. It had been found that different batches of zircon bars apparently fired under identical conditions had dissociated to different extents. The evidence for correlating creep resistance with zircon dissociation is, of course, not conclusive, but some support for the general view is available. Re-association can be made to occur by refiring the material and holding it at a lower temperature. Alternatively dissociation can be prevented by firing to a lower temperature. On the basis of results obtained, a maximum of 1700°C for the firing temperature was chosen. Further tests, including creep, indicated that zircon–chromic oxide fired to 1700°C will probably satisfy the unexacting requirements of a single-stage turbine.

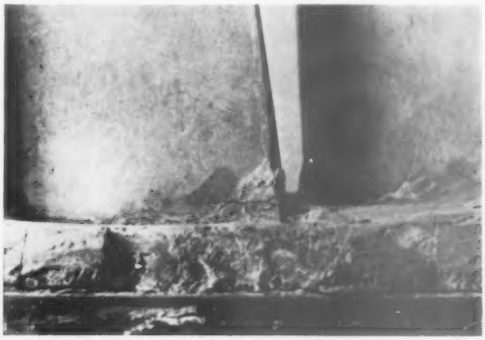


Figure 2. Typical cracks near blade roots

Figure 1 (CENTRE) shows a hollow and a solid blade made from zircon–chromic oxide.

Silicon Carbide—This too is a very promising material. Commercial silicon carbide bodies are usually made by bonding silicon carbide grain with clay. This however, adversely affects the properties of the pure material. More recently a new form of self-bonded silicon carbide, termed ' IC ', has been developed[3], which is more resistant to oxidation and distortion than the clay-bonded carbides and it combines good thermal conductivity, good high-temperature strength, and good creep resistance. Blades made from IC silicon carbide are shown on the right in *Figure 1.*

Metal–Ceramic Bodies—It is obviously desirable to combine in one material the high temperature strength of ceramics with the resistance of metals to thermal and mechanical shocks. Such materials known as metal–ceramics or ' cermets ' require further investigation. Cermets such as cobalt-bonded titanium carbide have been considered as possible blade materials. Tests however, were confined to the metal–metal oxide systems, of which a chromium–debased alumina body was chosen as being the one giving the best combination of ease of fabrication and desirable physical properties.

Testing of Blade Materials

The test rig consisted of a cascade of five blades mounted at the outlet of an oil-fired combustion chamber which discharged into the throat of a water-cooled diffuser. The gas temperature was measured by two refractory sheathed platinum/platinum–rhodium thermocouples just upstream of the blades, and there was also provision for measuring the pressure drop across the cascade. The blade cascade could be removed without disturbing the rest of the rig. Five blades were cemented in a metal channel piece between the refractory holding pieces.

First Series of Rig Tests

The first four cascades to be tested in the rig were all solid blades, slip-cast in zircon–chromic oxide,

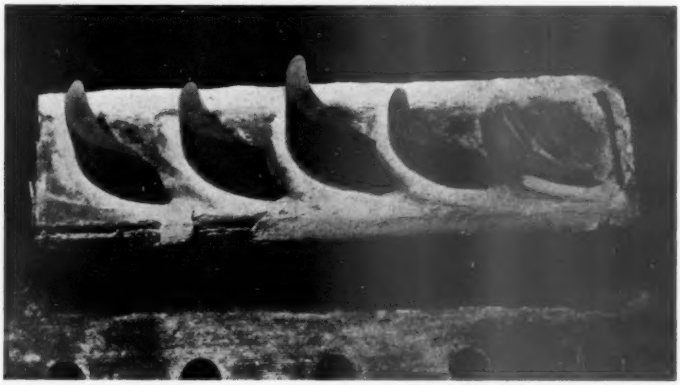


Figure 3. Fractured solid blades

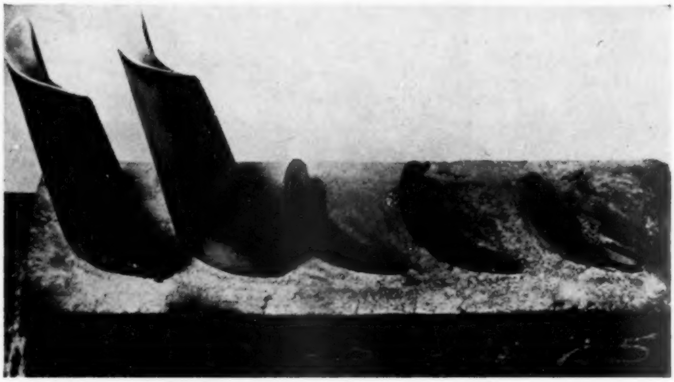


Figure 4. Fractured hollow blades

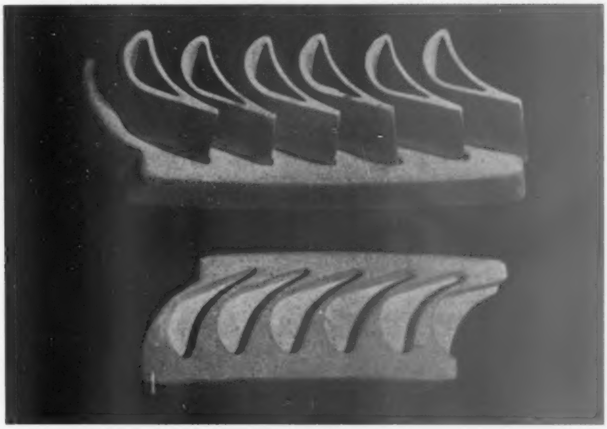


Figure 5. Blade segment partly dismantled

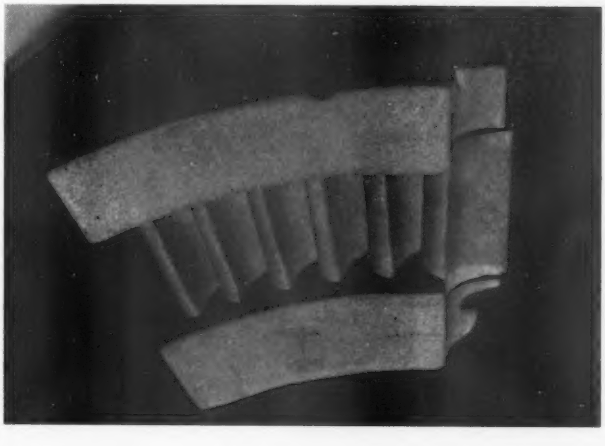


Figure 6. Blade segment assembled

with the roots gripped firmly between the packing pieces. The method of test was to light up as gently as possible and slowly increase the air and fuel flows until a temperature of 1205°C and a stress of 1000 lb/sq. in. were reached. Thereafter conditions would be held steady for several hours, or in certain cases a thermal shock was administered by abruptly reducing the fuel so that the temperature fell to about 705°C. On some occasions the stress was increased to 2000 lb/sq. in., maintaining the same temperature of 1205°C. All the blades fractured after a few hours running under such conditions, in several cases without any deliberate thermal shock being administered. The fractures were usually close to the root; typical instances of such fractures are shown in *Figures 2* and *3*.

It can be shown theoretically that the stresses set up in a hollow blade under conditions of changing gas temperature are proportional to the wall thickness, the wall thickness of a solid blade for purposes of comparison being taken as half the maximum thickness. The blades were fracturing under aerodynamic stresses much lower than the known strength of the material, and it seemed that the failure was due to thermally induced stresses. The next blades to be tested were therefore made hollow with a wall thickness of about one sixteenth of an inch, but they too failed even more rapidly than the solid ones. *Figure 4* shows a cascade of such blades, three fractured blades and two relatively intact.

On further consideration it became evident that the exposed portion of the blades would heat up more rapidly than the root, thus giving rise to exceptionally heavy stresses at the junction of the two portions, and this effect would be accentuated in the case of hollow blades. Therefore when a clearance was established, of the order of about 0·025 in., between the roots and the packing pieces, the first failure occurred after six and a half hours. After a total of about 19 hours running, two blades still survived the tests, which included eight starts and stops and four deliberate thermal shocks from 1205°C to about 705°C. This was a great improvement, and showed the need for supporting the blades loosely between the holding pieces.

Second Series of Rig Tests

This series consisted of tests on an experimental single-stage turbine with liquid-cooled rotor and refractory stator blades, to run at 1205°C. Being a single-stage machine, the fixed blades could be supported at both ends, thus greatly reducing the stresses and generally simplifying the problem of blade mounting. The mounting that was devised is shown in *Figures 5* and *6*. The blades are of prismatic form and both ends fit loosely into suitably shaped recesses in the outer and inner holding pieces. Twelve of them fit together to form a complete ring.

The next step was to test cascades of such blades in the rig. The water cooled diffuser could not be used in this case, and the test section had to be of segmented shape. The blade cascade was arranged to exhaust to atmosphere through a simple refractory-lined duct, and the rig, with the exhaust duct removed, is shown in *Figure 7*.

Modified Test Rig

The test procedure was much the same as before. Owing to the reduced velocity and pressure of the gas, the heat transfer coefficients under thermal shock conditions were considerably lower than would occur in an engine. To make up for this temperature difference, the normal practice was to reduce abruptly the gas temperature from 1205°C to about 705°C.

The first four or five cascades of blades to be tested were all of zircon–chromic oxide. The best of these blades showed themselves capable of withstanding numerous thermal shocks without visible damage. Others, developed cracks, usually in a longitudinal direction, as shown in *Figure 8*. It was discovered that the properties of this material were dependent to a considerable degree on the exact nature of the firing treatment. Once the correct firing conditions were determined and arrangements made to control them strictly, blades could be made that were apparently unaffected by twelve or more thermal shocks. Several such blades were run for prolonged periods at a steady temperature of 1205°C, and after periods varying from 80 to 150 hours they began to yield to the stress and eventually developed cracks as shown in *Figure 9*.

All these failures were considered to be due to what was essentially a form of creep. This was borne out by laboratory tests.

During the tests described above, the holding pieces had suffered much more severely than the blades themselves. The webs between the recesses

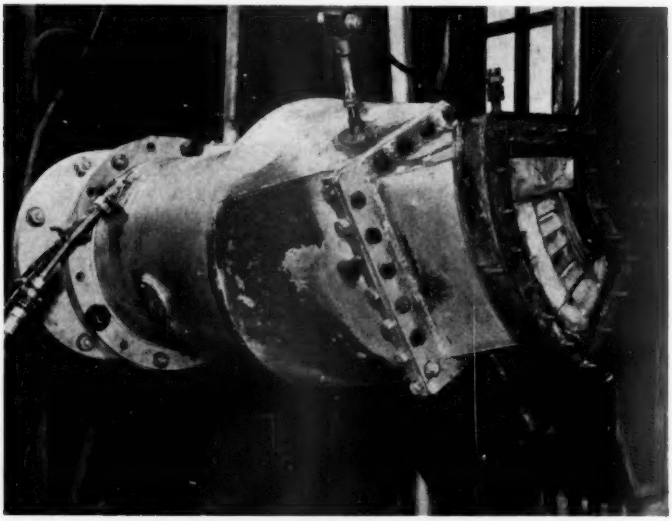
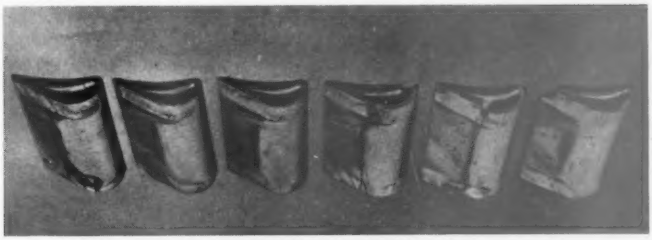


Figure 7. Test rig



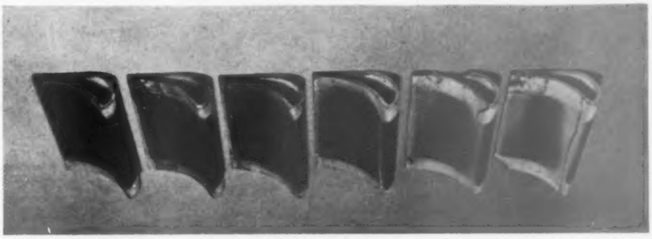


Figure 8. Longitudinal cracks after 19 hours running and ten thermal shocks

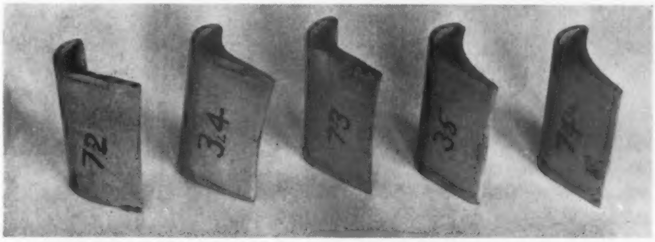


Figure 9. Zircon–chromic oxide blades showing distortion due to creep

cracked very easily and the overhung portion of the inner holding piece tended to break off. These holding pieces had been made of zircon, and it was decided to change to one of the proprietary materials[3] which was more resistant to thermal shock. At the same time a reconsideration of the design showed that the pitch/chord ratio of the nozzle blades could be increased somewhat without sacrifice of efficiency, and the total number of blades was therefore reduced from 72 to 60, *i.e.*, five blades per holding piece instead of six. The depth of the recesses was also reduced, and the combined effect of these modifications was to give a more robust holding piece with thicker webs. The new type of holding blade gave no trouble throughout the remainder of the rig tests.

A further material to be tried was a metal–ceramic body consisting of 75 per cent chromium and 25 per cent debased alumina. These blades showed themselves entirely resistant to the worst thermal shock that could be applied, but very weak in creep. At 1205°C they distorted appreciably in two or three hours, and even at 750°C they showed appreciable distortion in 15 or 20 hours, as will be seen from *Figure 10*.

Several self-bonded silicon carbide blades were tested and these proved themselves immune to the worst thermal shock that could be administered. One set was given sixteen shocks without any sign of damage appearing, and the test was subsequently carried on at a steady temperature of 1205°C for a total running time of 363 hours. The only damage suffered by these blades was a certain roughening of the concave faces, as shown in *Figure 11*. Analysis of this glassy material, which had evidently been molten at the working temperature, showed it to contain elements not present in the original blade, and it is believed to have come from inorganic material present in the combustion gases, not necessarily from the fuel.

Turbine Tests

The first of blades to be tested were the zircon–chromic oxides ones. It was considered advisable to test them first without the rotor in position and the turbine was therefore assembled in this condition. *Figure 12* shows the nozzle blade ring assembled in the turbine casing. The complete nozzle ring was then run for about 43 hours, including 11 hours at maximum temperature. Two of the blades were found to be cracked after a short period of running and two more after the completion of the above period. The remainder of the blades were in excellent condition.

The most serious damage was to the holding pieces, particularly the inner ones which are not supported on the downstream side. The material originally used was chosen entirely for its thermal shock resistance and it was evident that something with a greater strength would be required.

Conclusions

It is too early to make dogmatic statements on the outcome of the investigation into the use of non-metallic materials for stator blades. It is probable, however, that the development of metallic materials with good strength at high temperature is approach-

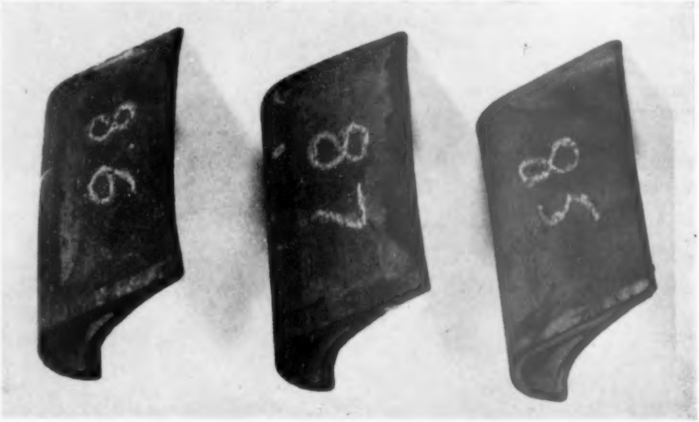


Figure 10. Chromium–alumina blades showing distortion

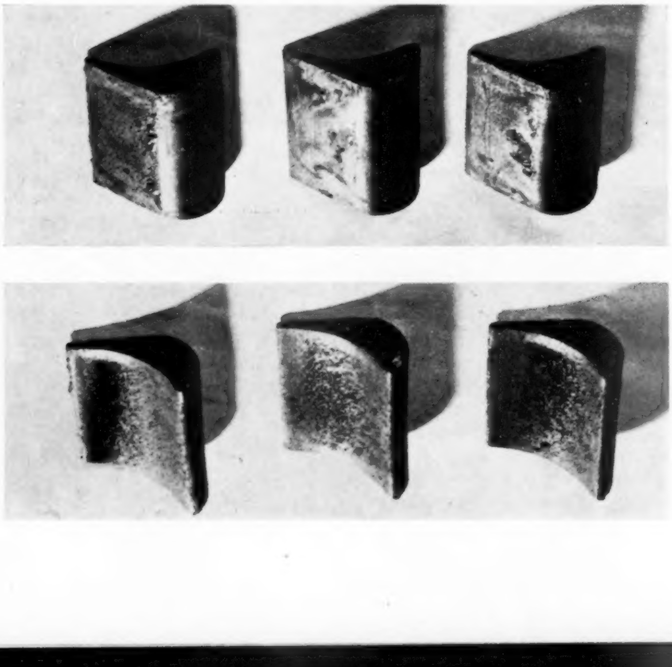


Figure 11. Silicon carbide blades after running for 363 hours

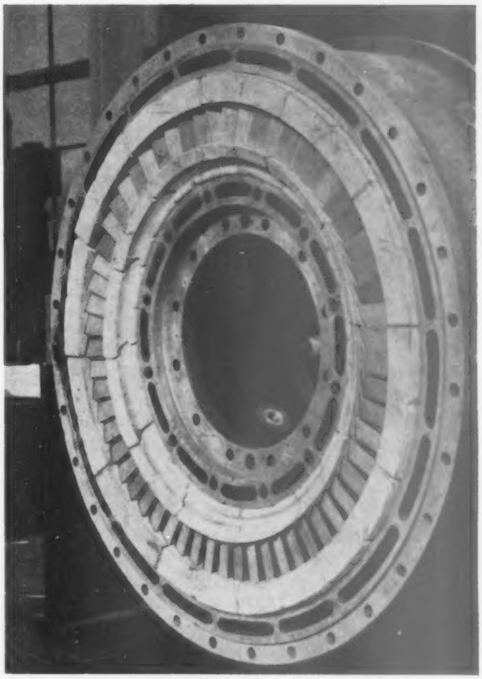


Figure 12. Blades assembled in turbine casing

ing its limit, and if gas turbine operating temperatures are to be raised to the same thermal efficiency level as a diesel engine, alternative types of blading must be sought. The choice lies between metallic blades, cooled in some manner or other, or non-metallic blades operating at the same temperature as the gas.

In turbine rotors the centrifugal field imposes high tensile stresses on the blades but also enables them to be cooled conveniently by the thermosiphon method, and in this case, therefore, the choice is likely to be in favour of cooled metal blades. Things are far otherwise in the fixed blading, however, where the stresses can be kept a good deal lower. Intensive investigation, has revealed the difficulty and complication of cooling a ring of nozzle blades effectively, and this is an obvious field for the application of non-metallic materials. There is an inherent difficulty in the relatively high brittleness of these materials, and this must be faced. An adequate factor of safety must be built with the design, and the designers, long accustomed to metals, must adapt their thinking to utilize the available materials, which at the best are somewhat akin to cast iron (at room temperature) in their behaviour.

The work described in this essay has indicated that stator blades can be made in a ceramic material (zircon) but that their factor of safety is low. Metal ceramic blades are more promising, but there is a nice balance between good creep resistance and good high temperature strength on the one hand, and good thermal shock resistance on the other.

More promising are the other materials already discussed, such as the newer forms of silicon carbide. These have excellent high temperature strength, good resistance to oxidation, thermal shock and creep. Fabrication of these bodies still presents difficulties; but here also, further improvement can be expected.

In tests to date the holding pieces have given, if anything, more trouble than the blades themselves, and the prevention of gas leakage is another serious problem which must be solved if progress is to continue. The answer here may lie in the use of metallic mountings—possibly protected with a coating of refractory—which could be relatively easily cooled as only a small amount of surface is exposed to the hot gas stream.

Although much work remains to be done, a promising start has been made and as the incentive is great, the future can be faced with considerable confidence.

The author wishes to thank the Editor of the Transactions of the North East Coast Institution of Engineers and Shipbuilders[3] *for permission to reproduce the illustrations with the exception of Figure 1. The author acknowledges gratefully the assistance rendered to this work by Mr E. S. White, of The Morgan Crucible Co. Ltd, and to Mr L. R. Barrett, of the Imperial College of Science and Technology for his guidance.*

References

[1] CARRUTHERS, T. G. and ROBERTS, A. L. *High-Temperature Steels and Alloys for Gas Turbines* p. 268 (Special Report No. 43). London: Iron and Steel Institute, 1952

[2] ROBERTS, J. R. and WATT, W. British Ceramic Society Report (October, 1950) *R.A.E. Tech. Note (Met) 46*

[3] BLAKELEY, T. H. and DARLING, R. F. *The Development of Refractory Nozzle Blades for Use in High-Temperature Gas Turbines* Paper to North East Coast Institution of Engineers and Shipbuilders (Feb., 1957)

THE DESIGN, MANUFACTURE AND TESTING OF INFRA-RED PHOTOCELLS

D. H. ROBERTS

The Plessey Co. Ltd, Caswell Research Laboratories, Towcester

This is the ninth article in the series on the technology of semiconductors which has been published at monthly intervals from October 1958 to May 1959. The present article considers photoelectric devices. It will be followed by three others dealing with other types of devices, intermetallic compounds with particular reference to bismuth telluride and organic semiconductors.

THE INVESTIGATION of the photosensitive properties of semiconductors preceded the discovery of the transistor by a great many years. Edmonde Becquerel discovered the photovoltaic effect in 1839, and Willoughby Smith was investigating photoconductivity in selenium in 1873. The first thallium sulphide cell was made by T. W. Case in 1920, and following on from that came the development of lead sulphide infra-red detectors in the early stages of the second world war—the most rapid development taking place in Germany. However, apart from the work of Gudden and Pohl in the 1920s and 1930s most of this work was of an applied nature, and for a complete understanding of the photovoltaic and photoconductive processes it has been necessary to await the stimulus given to solid state research by the discovery of the transistor at the Bell Telephone Laboratories in 1948.

Since that date there has been considerable work on the photo-effects in single crystal bodies of materials such as silicon, germanium, and indium antimonide, and it has been gratifying to find with these materials that we can accurately predict the observed behaviour—in complete contrast to the case of evaporated films of the lead sulphide group.

There are four types of quantum detectors—as opposed to energy detectors such as bolometers—but one of these, the photoemissive cathode does not strictly rely on the semiconducting nature of the cathode, so it will not be considered any further. The other three types are photoconductive, photovoltaic and photoelectromagnetic. These three all require quanta of energy greater than a certain threshold, this being the energy required to produce an electron–hole pair, or a single carrier in the extrinsic case. The subsequent behaviour of these excess carriers depends on the mode of operation.

Operation of Photocells

In the class of detectors known as photoconductors, the excess carriers produced by the incident radiation are observed as an increase in the conductivity of the specimen. It can be shown[1] that the signal voltage V_s from an intrinsic photoconductor is given by*

$$V_s = \frac{V_c \overline{Q}_0 \tau (1 - e^{aZ})}{\bar{n} Z (1 + \omega^2 \tau^2)^{\frac{1}{2}}} \qquad \ldots.(1)$$

where V_c is the bias voltage, $\overline{Q}_0$ the mean quantum flux density at the surface, τ the decay constant of excess carriers and a the absorption coefficient of the radiation. Z is the thickness, $\bar{n}$ the equilibrium carrier density and ω the angular frequency of modulation. (The last is equivalent to $2\pi f$, where f is the modulation frequency.) This equation defines the responsivity, usually given in volts per watt. It can be seen that given suitable matching, carrier mobility is unimportant but that the responsivity is $\propto \tau/\bar{n}Z$. In the various manufacturing techniques this fact together with the need to maintain a low value of S, the surface recombination velocity, is of paramount importance.

The Photovoltaic Detector

Any minority carriers, produced by the incident radiation, that reach the junction by diffusion are extracted by the local electrostatic field to become majority carriers. In the usual short circuit or matched load conditions a current flows through the load to preserve charge neutrality in the p- and n-regions. (The photovoltaic effect is shown in *Figure 1a* and *1b*.)

The short circuit current i_s is given by

$$i_s = Ae\overline{Q}_0\eta \qquad \ldots.(2)$$

where η is the collection efficiency and is a function

* A list of all the symbols used will be found in *Appendix I* on page 264

of the diffusion length, the decay constant of the excess carriers, the surface recombination velocity S, the junction depth and the absorption coefficient for the radiation. In regions of high absorption coefficient, η is of the order of 0·65 if S is assumed negligible and the junction depth equals the diffusion length of electrons in p-type material. These are then the basic design parameters: to maintain a low surface recombination velocity and control the junction depth.

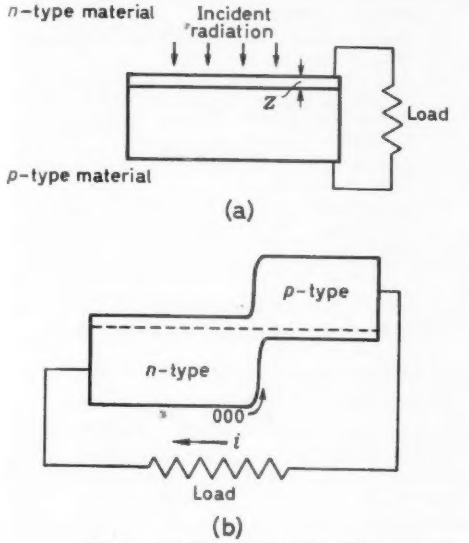


Figure 1. The photovoltaic effect

Photoelectromagnetic Effect

The photoelectromagnetic effect (see *Figure 2*) is the most recent mode to have been used in a practical photocell[2]. Non-penetrating radiation produces a concentration gradient of free carriers and hence a diffusion current of electrons and holes from front to back. If a magnetic field is now applied perpendicular to the plane of this diffusion, the positive and negative charge carriers are deflected in opposite ways, thus producing a voltage difference between the ends of the semiconductor.

If one assumes that the surface recombination velocity S at front surface is negligible, S at the back tends to infinity and the thickness of the specimen will be about the same as the diffusion length. One then obtains the maximum photocurrent i_{sc},

$$i_{sc} = \frac{A\overline{Q}_0 eB}{X}(L_{De}\mu_e + L_{Dh}\mu_h) \qquad \ldots.(3)$$

where L_D is the electron or hole diffusion length, μ the electron or hole mobility, A the receiving area of the detector, B the magnetic flux and X is the inter-electrode distance. And again, these conditions set our design parameters, it being necessary to control the surface recombination velocity both at the back and the front, and specimen thickness (apart from control of the material properties, lifetime, impurity concentration, carrier mobilities).

It is worth noting one important difference between the photoelectromagnetic and photoconductive modes, namely the dependence on thickness. In the first case specimen thickness should be about equal to the diffusion depth whereas in the second case the thickness should be nearly equal to $1/\alpha$, where α is the absorption coefficient for the radiation. (Some of the advantages of the photoconductive and photoelectromagnetic devices are compared in *Table 1*.)

Noise

Apart from the response of a photocell to a given amount of radiation its practical performance depends on the noise level in its operating condition. There are four types of noise occurring in photocells; more complete discussions of noise in semiconductors are given by R. E. BURGESS[3] and B. L. H. WILSON[4].

The first, known as the *Johnson noise*, is produced by the voltage fluctuations across any two terminal impedance Z in thermal equilibrium at absolute temperature T. At normal frequencies and temperatures the noise is given by

$$V_{JN}^2 = 4kT.\mathcal{R}(Z).\Delta f \qquad \ldots.(4)$$

where k is Boltzmann's constant, and Δf is the band width, and the Johnson noise can obviously be reduced by lowering the temperature, the detector impedance, and the bandwidth.

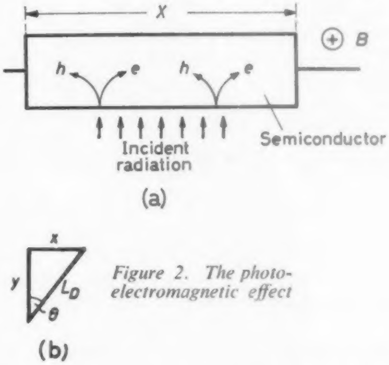


Figure 2. The photoelectromagnetic effect

There is excess noise, called *Shot noise*, in a semiconductor carrying a current compared with a

Table 1. Comparison of advantages of photoconductors and photoelectromagnetic detectors using indium antimonide

Photoconductors	*Photoelectromagnetic detectors*
(*i*) Superior spectral response—edge occurs at slightly longer wavelength	(*i*) Requires no bias supply—important in transistorized equipment
(*ii*) Less bulk and weight since no magnet is required	(*ii*) Easier to manufacture because of the thickness effect discussed earlier
	(*iii*) Likely to yield the more sensitive detector—because of greater responsivity of the photoelectromagnetic effect and lower noise level due to absence of bias current
	(*iv*) Less dependent on lifetime and surface recombination velocity
	(*v*) Easier to prepare suitable starting material, *p*-type doped rather than high purity intrinsic being used

metal and this is due to the fact that the excitation and recombination processes governing the number of free charge carriers are randomly occurring discrete events. (Burgess[3,5] gives a full treatment of shot noise.) Provided the lifetime of excess carriers is very much less than the transit time, it can be shown that for an intrinsic semiconductor, shot noise can be described by

$$\overline{i^2_{SN}} = 2eI.(\tau/\tau').\Delta f \qquad \ldots.(5)$$

where I is the d.c. current bias, τ the decay constant of excess carriers, τ', the transit time and Δf the bandwidth.

It is found that there exists a further type of current dependent noise, known as *flicker noise*. The origin and detailed mechanism of this type of noise are very much open to speculation[6]. Flicker noise may be described by the equation

$$\overline{i^2_{FN}} = E'I^u f^{-w}.\Delta f \qquad \ldots.(6)$$

where u and w are constants and approximately equal to two and one respectively. Some data on the dependence of E', and u on the properties of the surface and the presence of intercrystalline potential barriers was given in a paper by D. H. Roberts and B. L. H. Wilson at the symposium on *Noise in Fixed Resistors*, London, March 1958.

Photon noise is in a way analogous to shot noise, since it is due to the quantum nature of the incident radiation. In a photoconductor

$$\overline{i^2_{PN}} = 2C_2eI_0.\Delta f \qquad \ldots.(7)$$

where I_0 is the d.c. photocurrent produced by $\overline{Q}_0$ and C_2 is a constant. Photon noise in the signal is the ultimate noise in any detector, but, in room temperature detectors this is unlikely to be observed compared with the other three types. Photon noise is also generated by the ambient radiation particularly important in long wavelength detectors. The increase in sensitivity of lead sulphide and lead telluride cells on cooling their surroundings has been given by T. S. Moss[7] as evidence that they are photon noise limited.

Manufacturing Techniques: Photoconductive

Very sensitive visible radiation detectors made of cadmium sulphide have been made using powder techniques, with either a painting system with polycrystalline powder[8] or sintering a film of amorphous powder[9]. Infra-red photoconductors however have been made principally by sublimation[10], chemical deposition[11] and the use of lapped and etched single crystal filaments[12]. These will now be considered in more detail below.

Evaporated Film Method

The evaporated film method has been widely used with the lead sulphide group of photoconductors. As an example lead selenide films have been prepared by evaporating starting material of the form $Pb_5Se_3O_3$. This material was prepared by mixing the appropriate amounts of lead, lead oxide PbO, and selenium, sealed in a pyrex tube at a pressure of less than 10^{-5} mm, fired at 450°C and the final product was crushed and sieved. A few milligrams of this starting material were then placed in a pyrex glass cell blank, with bubble window and aquadag–tungsten electrodes, which was sealed to a vacuum system and evacuated to better than 10^{-5} mm of mercury. The actual sublimation process was performed by flame-heating the base of the cell blank, causing a film of lead selenide to be deposited on the window between the electrodes.

This lead selenide film then required sensitizing by heat treatment in air, the choice of time, temperature, and air pressure being determined by the characteristics of the film as prepared and the whims of the operator! During the sensitization, the conductance and photoconductivity of the film under standard illumination would be measured in a simple d.c. circuit. The variation of these and other parameters, thermoelectric power and carrier lifetime, with sensitizing treatment is given in a recent paper[13]. At the appropriate stage the cell blank would be evacuated and sealed off ready for use.

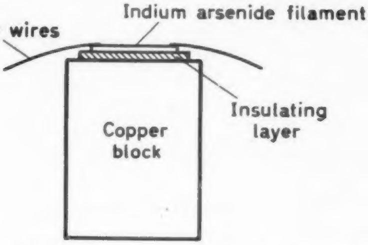


Figure 3. Indium arsenide photoconductive cell

Chemical Deposition

Chemical deposition methods have been used and again lead selenide will be used to give a specific example, although this method has been most widely used on lead sulphide by B.T.H. Ltd in this country and Eastman Kodak, among others, in the United States. Films of lead selenide are reproduced from the colloidal phase in the reaction between aqueous solutions of a lead salt and seleno-urea in the presence of a base. A glass substrate having been degreased and washed was rotated in a solution containing lead acetate and seleno-urea, hydrazine hydrate being used as the base additive to control the reaction[11]. Having deposited a film of the required thickness the substrate was removed from the solution, washed, and dried. A suitable electrode system could then be painted on the film using aquadag or a colloidal gold suspension. Workers using lead sulphide have subjected their chemically deposited layers to a tempering or sensitizing treatment by baking the deposited films in air at atmospheric pressure—sometimes for quite long periods.

Solid Filaments Technique

Whatever failings the solid filaments technique may have, it does possess one major virtue. It has been possible with materials such as indium antimonide[12] and indium arsenide to construct photocells whose performance could be understood and whose sensitivity could be predicted theoretically with a fair degree of accuracy. The method described below is the most recent application of this technique to indium arsenide.

A zone refined ingot of indium arsenide was cut into slices, these slices being lapped to a thickness of about 0·1 mm and then diced to give filaments 3×1 mm which were suitable for mounting and etching. In order that the detector shall not be dissipation limited[1] it is necessary to provide an efficient heat sink. This was done as shown in *Figure 3* by mounting the indium arsenide filament on a copper block with a thin insulating layer isolating the filament electrically whilst permitting good thermal contact. Wire contacts were soldered on to the filament to connect it across a co-axial socket. After mounting, the region around the contacts was masked with a hard setting resin or wax and then the specimen was etched to reduce the thickness toward the ideal of a few microns and to reduce the surface recombination velocity.

Measurements performed on this type of detector indicated that in the near future it will compete with lead sulphide cells for use in the 1 to 3·5 micron region.

Manufacturing Techniques: Photovoltaic Detector

The photovoltaic detector has certain virtues (and disadvantages) compared with the photoconductor which are briefly summarized below.

(*i*) The photovoltaic effect is less dependent on lifetime and surface recombination velocity.

(*ii*) It is not necessary to reduce the overall thickness to such mechanically difficult levels.

(*iii*) The devices are not so easily de-sensitized by an increase in temperature or background illumination.

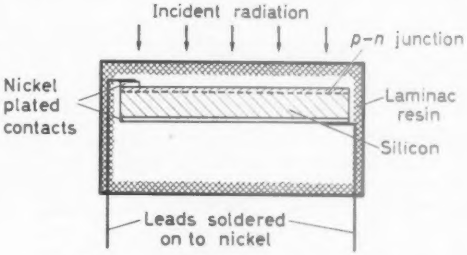


Figure 4. Silcon photovoltaic detector

The disadvantages are two fold: more advanced materials technology is required to produce a p–n junction and for long wavelength detectors, *e.g.*, indium antimonide, it is necessary to dope so heavily to obtain a p–n junction at room temperature that the carrier lifetime and mobility are degraded very considerably.

As an example, the method of making a silicon photocell will be described. This is basically the same technique as is used to construct a solar battery for the conversion of the sun's radiation into electrical energy. Assuming one started with a single crystal of p-type silicon it is necessary to diffuse an n-type impurity, *i.e.*, phosphorus into a slice of silicon to obtain a p–n junction. In order to make a device however it is necessary to make electrical contacts to both the p- and n-regions. A very convenient way of doing this is to use what is known as the electroless nickel plating technique, but in practice it is difficult to make an ohmic contact to high resistivity silicon, so that a p^{+++} region is produced in the base region on to which the nickel can be plated.

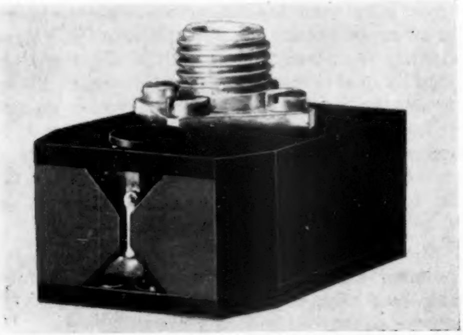


Figure 5. Infra-red detector (by courtesy of Technical Ceramics Limited)

The procedure, therefore, is as follows. Boron is diffused into the slice[14] and the boron doped region is removed from one side by grinding. Phosphorous is then diffused into the slice and the resulting oxide film on the surface removed by hydrofluoric acid. After nickel plating the whole slice the active regions were exposed by etching in nitric acid, and the junction was etched in a mixture containing equal parts of the two acids, having masked the remainder of the structure. After washing and degreasing the nickel plating was tinned, using a dip soldering technique; nickel tape leads were then soldered on and the cell potted in a hard-setting resin. The final structure is illustrated in *Figure 4*.

Manufacturing Techniques: Photoelectromagnetic Detector

The methods used for the manufacture of photoelectromagnetic detectors have somewhat limited application, it being unlikely that it will be used with any material other than indium antimonide. A comparison between this and the photoconductive type of detector is given in *Table 1*.

As already mentioned the first stage in the construction of indium antimonide photoelectromagnetic detector is to cut and lap the single crystal to produce filaments 3 × 1 × 0·1 mm. These are mounted on an insulator with contacts attached, as there is in this case no need to consider dissipation. Once mounted the specimen was etched in, for example, a mixture of hydrofluoric acid and hydrogen peroxide. In early development the procedure was successively to etch and measure the performance until the optimum thickness is reached and once a crystal had been characterized a large number of detectors could be made by etching for a standard time. The specimen was then mounted in a magnet, a final assembly being shown in *Figure 5*. The magnets used are made of Alcomax III with mild steel yoke and permandur pole pieces shaped to give a field of about 10,000 gauss in a gap of 1·5 mm.

Methods of Testing Photocells

The spectral response is the variation of responsivity with wavelength, and in principle it is only necessary to measure the ratio of input and output signals at various wavelength, the input, *i.e.*, radiation, signal usually being measured with a thermal detector of known responsivity such as a Schwartz radiation thermocouple. In practice there are many variations and many precautions must be taken. For the single beam system one must chop at the highest possible frequency to differentiate between photoconductive and bolometric effects and also chop at the input slit of a monochromator to avoid spurious signals. For double beam systems the exact method used depends on the equipment available. Some of the variables are beam splitting, measurement of signal ratio, radiation source and area of thermal detector. *Beam splitting* involves combining with chopping at the output slit or placing a mirror in half the optical field or a half-silvered mirror in whole of the optical field. The last of these suffers the least disadvantages and is the most reliable.

Measurement of signal ratio is carried out by measuring two signals individually and computing the ratio at fixed points, by applying both signals (rectified) to a potentiometer recorder connected as a ratiometer or by using the reference signal to control the source output, or slit width, to give constant energy at all wavelengths on the detector under test. If the last method is used, a tungsten filament lamp with a suitable window to transmit infra-red, can be used as a *radiation source*. Otherwise a Nernst filament is suitable provided care is taken to shield the filament from draughts. To determine the *area of the thermal detector* care must be taken, in a double beam system, to ensure that all the radiation signal is incident on the reference detector. This is easier if the detector area is comparable with the photocell under test.

Finally it is worth noting that such a technique will give a curve of responsivity in volts per watt versus wavelength, whereas from the semiconductor point of view volts per number of quanta is more meaningful and the simple conversion is worthwhile.

Responsivity and Noise

To determine the responsivity and noise, and hence the equivalent noise power, the signal source used is 'a 200°C black body'. This consists of a temperature controlled furnace whose interior can be 'seen' by the detector through a series of baffles whose geometrical arrangement is such that the final aperture behaves as a perfect black body at 200°C, thus allowing the radiant flux at a standard distance to be calculated quite easily. The radiation output is generally modulated at 800 c/s and the signal from the detector placed at a standard distance, 20 cm, is fed into a tuned amplifier with a bandwidth of 50 c/s. The noise and signal levels are then measured, in the photoconductive case as a function of bias current. The ratio of input and output signals, at the detector, gives the responsivity in volts per watt. This together with the noise level allows the equivalent noise power, defined as minimum incident energy to give a signal-to-noise ratio of unity when measured in unit bandwidth, to be calculated.

With photoconductive cells where flicker noise is suspected it is essential to check the current and frequency dependence of the noise voltage as these will determine the choice of operating conditions.

Response Time

It is necessary to have a pulse of radiation or a modulated sine wave of variable and high frequency to determine the response time. The simplest method is probably to use a triggered spark discharge as source and display the signal on a cathode ray tube with calibrated time base, the response time, τ, being defined as the time for the signal to fall to $1/e$ of its initial value. Care has to be taken to guard against spurious time constants. If a fast sine wave modulation is used there are two possible approaches: to measure the frequency dependence of responsivity R

$$R = \frac{R_0}{(1 + \omega^2\tau^2)^{\frac{1}{2}}} \qquad \ldots.(8)$$

or to measure the phase difference δ, between input and output signals

$$\delta = \tan^{-1}\omega\tau \qquad \ldots.(9)$$

and hence calculate τ from *(8)* or *(9)* where ω is the angular frequency.

It is worth mentioning that these same methods can be used to measure the carrier lifetime in a bulk semiconductor, for example, silicon, but it is necessary to take precautions. In the case of traps the effect of d.c. illumination on the shape of the decay curve must be observed. The short wavelengths are filtered out and again the effect on the decay curve due to surface recombination is observed in order to measure carrier lifetime. Lastly to measure the illumination level, it is necessary to standardize on signal level as τ can be intensity dependent.

Conclusions

In the last few years the range of available infra-red detectors has expanded quite considerably, as also has the understanding of their performance and its dependence on basic material properties. It is to be expected that the next few years will see an even greater expansion in the application of these photocells in industry. In particular the use of these devices for non-contact temperature measurement enables surface temperatures to be measured down to only a few degrees above ambient with, for example, the indium antimonide detector. Furthermore the combination of this type of detector with multilayer interference filters for the infra-red offers the possibility of making cheap, small, robust, sensitive instruments for gas analysis and chemical process control.

In a different field, the combination of photoconductive and electroluminiscent elements is already being heralded as opening new paths for the computor designer and time will possibly bring an 'optical' computor.

The author is grateful to the Plessey Company Ltd for permission to publish this article.

References

[1] ROBERTS, D. H. and WILSON, B. L. H. *Brit. J. appl. Phys.* (1958) **9**, 291

[2] HILSUM, C. and ROSS, I. M. *Nature, Lond.* (1957) **179**, 146

[3] BURGESS, R. E. *Brit. J. appl. Phys.* (1955) **16**, 385

[4] WILSON, B. L. H. *J. Brit. Instn. radio Engrs* (1958) **18**, 208

[5] BURGESS, R. E. *Proc. Phys. Soc. B* (1956) **69**, 1020

[6] VAN VLIET, K. M. *Proc. Inst. radio Engrs* (1958) **46**, 1004

[7] MOSS, T. S. *ibid.* (1955) **43**, 1869

[8] NICOLL, F. H. and KAZAN, B. *J. opt. Soc. Amer.* (1955) **45**, 647

[9] THOMSEN, S. M. and BUBE, R. H. *Rev. sci. Instrum.* (1955) **26**, 664

[10] SOSNOWSKI, L., STARKIEWICZ, J. and SIMPSON, O. *Nature, Lond.* (1947) **159**, 881

[11] ROBERTS, D. H. and BAINES, J. E. *J. Phys. Chem. Solids* (1958) **6**, 184

[12] AVERY, D. G., GOODWIN, D. W. and RENNIE, A. E. *J. sci. Instrum.* (1957) **34**, 394

[13] ROBERTS, D. H. *J. Electron. & Control* (1958) **5**, 256

[14] FULLER, C. S. and DITZENBURGER, J. A. *J. appl. Phys.* (1956) **27**, 544

Appendix I will be found on page 264

ELECTRICITY GENERATION FROM MOVING GAS STREAMS

R. G. VOYSEY

Chief Scientists Division, Ministry of Power, Thames House South, Millbank, London, S.W.1

Considerable interest has been shown in thermal-electric generators with no moving parts other than a gas stream because it would be possible to operate such devices at very high temperatures which could well correspond to high thermodynamic efficiencies. The author points out that although the magnetodynamic form of ionic generator shows little promise, an electrodynamic generator rather like a van der Graaff machine but using ionized particles has potential applications.

FARADAY'S discovery of the magnetodynamic generation of electricity was prompted by Arago's observation of the deflection of a magnetic needle near a rotating copper disc. Arago claimed that the effect held with a rotating conductor of any sort, including liquids and gases. This was a bold claim since the resistivity, even of a Bunsen flame, is nearly 10^{12} times that of copper. Faraday reports[1] that Babbage and Herschel could not demonstrate Arago's effect with any non-metal except gas-coke and he himself could not obtain significant results with sulphuric acid or salt solutions. However he later obtained erratic readings of the flow of the 960 ft width of the Thames at Waterloo Bridge by measuring the voltage induced by the motion of the brackish water through the earth's field[1].

Since Faraday's time, and particularly after the clear formulation of the conception of ions and electrons, repeated patent proposals have been made for ionic generators using not metals but fluid, preferably gaseous, conductors to secure some special advantage of efficiency or applicability. For example, 1917 was a year of great interest in trans-Atlantic radio transmission and A. YOUNG[2] suggested using a fluid conductor device for the generation of low frequency radio waves. In the next few years patent applications for ionic generators appeared in several European countries[2–4]. These and subsequent inventors had little conception of the problems but were attracted by the apparent possibility of extremely compact thermal-electric generators having no moving parts other than a gas stream and susceptible to operation at very high temperatures, which might imply a correspondingly high thermodynamic efficiency.

The magnetodynamic form of ionic generator is probably unattractive but an electrodynamic form, rather like a van de Graaff machine but using ionized particles might be feasible if properly designed. While cold, van de Graaff type machines using dust to carry charge are well developed (early investigators were Armstrong and Faraday[7]) but there is little evidence of thermal-electric machines with ionic current carriers being made. If experimental work has been done it probably exposed two serious difficulties. These are the high thermodynamic cost of providing sufficient current carriers in the gas stream and the possible instability of these ionic generators to internal disturbances of various kinds. These difficulties are worth examination after considering the principles and the simplest forms which such devices might take because there is renewed interest in efficient light weight thermal-electric generators of all sorts and a growing appreciation of the subtler points of magneto-hydrodynamics which might encourage a practical ionic generator.

As a device employing no machinery, other than its auxiliary compressor or boiler, it invites comparison with the fuel cell, the thermocouple generator and the thermo-electron generator. The first two have been the subject of articles in RESEARCH[8,9] while the third is being actively studied in the United States[10–14]. Unlike these three, the voltage output of the ionic generator is not controlled internally by considerations of the material work functions and, in the electrodynamic form, prefers very high voltages controlled by external fields. In theory the ionic generator suffers from being a semi-mechanical heat engine and not truly a direct generator of electricity from heat. Its promise of high temperature operation makes this consideration far less important than practical difficulties. If feasible it should be far more compact than the other three.

Principles of Generation

The particular class of electrodynamic machines here called ionic generators collect the charge of gaseous ions, *i.e.*, charged carriers, propelled against electrostatic or magnetic fields, or combinations of the two, by a moving gas. The electrical output is maintained by a back pressure on the gas flow.

Such devices have been suggested, operating in reverse, as ion rockets[15] and also, working in reverse with liquid metal conductors, they are electromagnetic pumps.

Before considering physical forms of the device it is possible to make an estimate of the required degree of ionization of the gas for various voltages of generation.

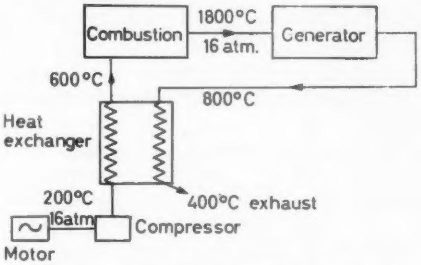


Figure 1. Heat engine cycle—ideal conditions only and not necessarily achievable in practice

Taking some specific thermodynamic cycle, for example the very speculative Joule cycle of *Figure 1*, a figure can be chosen for the desired isentropic heat drop in the device, in this case 240 C.H.U.s per pound. Irrespective of the machine's size or mass flow rate this leads immediately to a relation between the necessary ion population and the voltage of operation. *Table 1*, based simply on the conservation of energy, shows the tolerable scarcity ratio of ions, *i.e.*, the number of neutral molecules per ion which must not be exceeded to carry an adequate current at various voltages for 100 per cent efficiency of generation. Even a simply ionized molecule, involving the detachment of one electron, requires the energy equivalent of about 20,000°C per molecule. *Table 1* shows that on the most optimistic expectation of operation, a voltage near 30 kilovolts implies a loss of energy in ions at the exhaust equivalent to 20°C when averaged over all the molecules. At 600 volts the loss would average 1000°C and compare with the isentropic temperature drop assumed in the cycle of *Figure 1*. A further loss is entailed in providing electrons at the output electrode to neutralize the arriving positive ions.

In small scale application efficiency is often less important than extremely light weight and simplicity, but quite high voltages would still be demanded. This requirement almost certainly disposes of the simplest form of ionic generator suggested (see for example *Figure 2*) but is less damaging to the device suggested in *Figure 3*. These forms are described under the next heading.

Elementary Forms of Ionic Generators

The conception of *Figure 2* is the favourite of inventors. The gases in passing through a suitable expanding nozzle also travel through a transverse magnetic field. Under the combined action of the magnetic field and the gas velocity any charged particles in the gas will try to traverse orbits having a circular motion plus a steady drift in a direction at right angles to both the gas flow and the magnetic field. They will also suffer repeated collisions. With a possible and proper choice of design conditions the sideways drift preponderates over the circular motion and some carriers are imagined to be collected finally by side electrodes. The theoretical maximum voltage generated is simply the number of magnetic lines cut per second, divided by 10^8. Even with fields of 80,000 lines per square inch and gas velocities of 1000 feet per second, the potential generated can only be of the order of ten volts per inch width of nozzle.

A number of considerations limit the nozzle width so, if the device is to be efficient, it must produce extremely large currents to harness the very high rate of energy conversion from heat to velocity which takes place in the nozzle. The large current in turn implies a high proportion of ions in the gas stream so that this idea might only be made to work at all with elaborate and energy-consuming ion injectors and collectors. Without them it would give only a tiny fraction of its theoretical output and correspondingly little resistance to gas passage through the nozzle.

Table 1

Neutral molecules per ion	*Required voltage* (in kV)	*Average temperature loss with 20,000°C ions* (in °C)
50,000	1515	0·4
20,000	606	1
10,000	303	2
5000	151·5	4
1000	30·3	20

Inventors are tempted to believe that some special circumstance, for example the use of some easily ionizable medium like caesium vapour can help. It is true that at the gas temperatures in which we are interested a small amount of alkali metals (or better, the alkaline earths) give a disproportionate increase in conductivity. This still leaves the conductivity very low indeed and the gas also contains charges of both signs which is useless for this device. These matters are discussed further under the heading concerned with the production of ions. It is worth repeating that the table is based on simple arguments of energy balance. The results are expressed more generally by the following relation: the required fractional concentration of ions is approximately 0·00043 *WM*/*V*. (*W* is the specific output of the gas in C.H.U.s per pound flow, *M* is the molecular weight of the substance and *V* is the voltage of generation.)

The second simple conception of an ionic generator is the electrodynamic one of *Figure 3*. Advantage is taken in this of the high dielectric strength of gases, even at high temperature, to work at very high voltage. The ions are propelled against the repulsion of the collecting electrode. Modern work (see reference 17 and its sources) has shown the general design requirements (particularly for the geometry and terminal arrangements) of normal van de Graaff machines and these need not be treated here. The gases in the ionic generator would flow at over ten times the highest speed of the charging belts in the normal van de Graaff. Elementary consideration shows that while the gas velocity may be high and the distance between starting and collecting electrodes can be made so small that the transit time might be of the order of ten microseconds, a useful charge carrier might—and would desirably—suffer at least a thousand accelerating collisions in transit. There is then the difficulty that an avalanche of ionization might set in with the production of a conductive path in the stream through which the electrical energy could arc back. This is part of the general problem of instability but before it is discussed further it will be advisable to clear away the simpler matter of suitable thermodynamic cycles.

Suitable Thermodynamic Cycles

The arguments behind *Table 1* do not prejudice the choice of medium. If any sort of ionic generator is ever made, some practical requirement such as the ionizing mechanism might restrict the medium but there is no criterion for limitation at the present stage. It is therefore possible to consider cycles using any condensible or non-condensible gas. As with most thermal-electric devices it is helpful in considering efficiencies to separate the ideal thermal cycle from the electrical process. With an ionic generator one can think of the appropriate cycle, say Rankine's for condensible media—water, mercury, sodium, caesium *etc*—or the Joule cycle (as in *Figure 1*) for gases like air, combustion gases, monatomic gases *etc*. With these two cycles the generator component is the expander-plus-dynamo of a steam or gas turbine cycle and these are well studied fields.

It would be out of place to review the extensive literature on these cycles and their available media. Briefly, if practically achievable, the ionic generator will derive great benefit from its lack of machinery and enjoyment of high temperature. The Joule cycle conditions of *Figure 1* suggest a target cycle efficiency of 83 per cent. A Rankine cycle using some suitable low pressure vapour, say mercury between 750°C and 120°C, can surpass 50 per cent efficiency. It is also possible to think of the generator as incorporated in the relevant stage of any reciprocating cycle, *e.g.*, in the Otto or Diesel cycle, with a further expansion of its imaginable embodiments.

Most inventors or promotors of an idea make the mistake of claiming overwhelming advantages which may very well be real in the long run but compromise the early limited success which might be achieved by avoidance of immediate problems of deficient knowledge and technique. The basic thermal cycles which an ionic generator might use are so efficient that it is advisable to exploit them to relax design difficulties elsewhere. Thus the device might be developed as an inefficient light weight device or only as a superposition unit, drawing a valuable few per cent of energy from a hot gas and exhausting to a more normal heat engine component like a boiler or turbine.

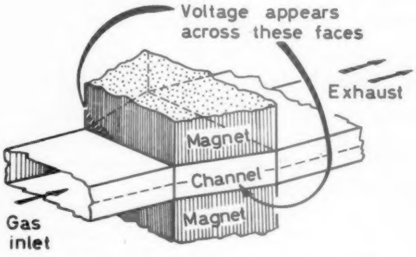


Figure 2. Magnetodynamic generator

It is difficult to imagine the device working consistently on other than ash-free fuels.

Production of Ions and the Possible Instability of the Device

It may have been noticed that the word ' plasma ' has been avoided so far. It suggests a more extreme condition of a conducting fluid in which most of the molecules are ionized at least singly, carriers of both sign are present, and the gas is in something near thermodynamic equilibrium at 15,000° to 20,000°C or above. None of these conditions is either necessary or desirable here unless the generator is imagined to derive its energy from low cost nuclear sources. Every re-inventor of the device thinks of radioactivity as an easy source of ionization but it is an ill matched source. Ionization needs collision and radioactive bombardment represents an extreme case of using particles of very many electron volts to liberate electrons frpm bonds of only a few electron volts energy.

At the other extreme inventors have hoped for sufficient ionization by gentler collisions at thermal velocities. A proportion of ions may be produced even in cool gases by chance collisions of exceptionally fast molecules but the concentration could not rise to useful values, even for kilovolt operation, until the gas temperature was above 2000°C.

While the curious behaviour of hot gases in a magnetic field has been noted for over a century, see, for example, Faraday's experiments[18], the physical conceptions for explaining their electrical behaviour were lacking till this century. Even now, while technology spreads towards the application of very high temperatures, it is still impossible to estimate the conditions of ionization in hot gases except under very restrictive assumptions. The first attempt, by Saha for estimating the conditions within a star, led to an equation which is discussed in a chapter on the ionization of flames[19].

The Saha equation at the outset assumes equal numbers of positive and negative ions and involves the product of these numbers to evaluate an equilibrium constant for their partial pressures. If a vanishing proportion of negative ions is assumed, the equation would justify the stability of a very large proportion of positive ions. In the ultimate, such a cloud would be self-repelling, ignoring any ' pinch ' due to motion, but the small electric pressure could easily be balanced by potential on the walls. A more obvious difficulty might be the encouragement of electron emission from both walls. It seems probable that this will always take place to some degree even if the wall is built up from a series of guard rings of graded potential and it implies a limitation of working temperature and the use of materials of very high work function for everything except the electrodes. It also discourages the scattering of alkalis or alkaline earths into the gas stream. It has already been said that the advantages of such additions are probably not real at the required conductivities. The device could not rely on thermal ionization alone and indeed the latter would, if significant, destroy its stability. The feasibility of operation rests essentially on economically producing a good supply of ions of one sign and then taking advantage of the ' freezing ' of electronic reactions at the comparatively low temperatures to complete the few microseconds of travel through the nozzle.

Considering the simple electrodynamic generator the charged particles can only proceed against the voltage gradient by replenishing their kinetic energy from uncharged neighbours in their decelerating path. A number of collisions approaching 1000 would meet the case without running more than a negligible risk of meeting sufficient very high energy particles to initiate avalanching. There is sufficient design freedom in the problem to keep down to this order of number of collisions.

In view of the restricted prospects of the magnetic conception of *Figure 2* there is no point in discussing the very complicated ion behaviour which should take place in it.

Even if the electrostatic form of *Figure 3* is thought promising, design is complicated by the interaction of the generated current with itself and the surroundings (as with armature reaction problems in rotating machinery or ' pinch ' *etc* in plasmas proper). While Maxwell's equations tidy up electromagnetic behaviour very nicely, it is difficult to reconstruct physical effects from these laws. We can now follow nuclear pioneers and take over their developed ideas of ' pinch ', ' magnetic mirrors ', ' magnetic bottles ' *etc* but these effects are very much modified in the generator's region of relatively low temperature, low ion density and mobility. Consideration suggests that with sensible proportioning of the nozzle there should be no fundamentally unsolvable problems of loss of ions or pinch. The more obscure electromagnetic effects may be exploitable. The 1958 Geneva Conference on Nuclear Energy contained many papers on the dynamics of plasmas proper but only a few of the papers are at all relevant here; a good general reference is the symposium[20].

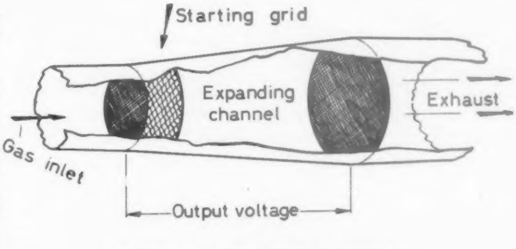


Figure 3. Electrodynamic generator

Conclusions

Despite repeated attempts to patent an ionic generator no one appears to have made any practical form of the device. This is not surprising in view of the daunting complexity of design and operation which underlies the superficial simplicity of its conception. Yet its basic thermal efficiency and its high power–weight ratio are so promising that it is surprising that so little experimental work has been done. Some exploratory measurement on ion densities was made by M. W. THRING and others at the University of Sheffield in 1958 with the support of the National Research and Development Corporation.

Much more work could be done on the electrostatic form which may have early application for small unattended power sources. If such developments could demonstrate the general feasibility of the ionic generators, the field of application is potentially very wide because of its simplicity, its good thermal cycle and small size.

The author is grateful to the N.R.D.C. for permission to mention their interest and work. Opinions and conjectures expressed are the author's own but he is very grateful to the Ministry of Power for the opportunity to make this study and permission to publish this note.

References

[1] FARADAY, J. *Experimental Researches in Electricity* (1838) **1**, 81, 130, 188

[2] YOUNG, A. *Brit. Pat.* 113,679 (first application 1917)

[3] PETERSEN, C. *Brit. Pat.* 122,173 (first application 1918)

[4] BERTHET, J. L. *Brit. Pat.* 214,223 (first application 1923, not finally accepted)

[5] DE KRAMOLIN, L. L. *Brit. Pat.* 414,851 (first application 1932)

[6] *U.S. Pat.* 2,210,918 (first application 1936)

[7] FARADAY, J. *Experimental Researches in Electricity* (1844) **2**, Section 25, 2075

[8] WATSON, G. H. *Research, Lond.* (1954) **7**, 34

[9] GOLDSMIDT, H. J. *ibid.* (1955) **8**, 172

[10] MEDICUS, G. and WEHNER, G. *J. appl. Phys.* (1951) **22**, 1389

[11] HALSOPOULOS, G. N. and KAYE, J. *Proc. Inst. radio Engrs* (1958) 1574

[12] WILSON, V. C. *J. appl. Phys.* (1959) **30**, 475

[13] HOUSTON, J. M. *ibid.*, p. 48

[14] WEBSTER, H. G. *ibid.*, p. 488

[15] VON ENGEL, A. *Nature, Lond.* (1959) **183**, 573

[16] TRUMP, S. G. *Sources of Electric Energy*, p. 9 (Conference on Energy Sources) New York: American Institute of Electrical Engineers, 1951

[17] FARADAY, J. *Experimental Researches in Electricity* (1855) **3**, 467

[18] GAYDON, A. G. and WOLFHARD, H. G. *Flames, their Structure, Radiation and Temperature* London: Chapman and Hall, 1953

[19] LANDSHOFF, R. K. M. (*Ed.*) *Magnetohydrodynamics—A Symposium* Stanford, California: Stanford University Press, 1957

[20] *2nd International Conference on the Peaceful Uses of Atomic Energy* (*Geneva, 1958*) New York: United Nations, 1959

Continued from page 259

The Design, Manufacture and Testing of Infra-Red Photocells

D. H. ROBERTS

Appendix I

For the convenience of the reader a complete alphabetical list of all the symbols used in the article is given below.

A	=	receiving area of detector
a	=	absorption coefficient for radiation
B	=	magnetic flux
C_2	=	constant describing fluctuations in photon noise
E'	=	constant describing flicker noise
e	=	electronic charge
f	=	frequency, modulation frequency
Δf	=	bandwidth
I	=	d.c. bias current
I_0	=	current corresponding to incident radiation flux Q_0
k	=	Boltzmann's constant
L_D	=	diffusion length
L_{De}, L_{Dh}	=	electron and hole diffusion length
μ_e, μ_h	=	electron and hole mobility
n	=	equilibrium electron density
Q_0	=	mean quantum flux density at surface
S	=	surface recombination velocity
T	=	absolute temperature
τ	=	decay constant of excess carriers
τ'	=	carrier transit time between electrodes
V_c	=	bias voltage
V_s	=	signal voltage
ω	=	angular frequency $= 2\pi f$
X	=	inter-electrode distance
Z	=	specimen thickness, junction depth
$\mathcal{Z}$	=	impedance

MICROWAVE AMPLIFICATION BY STIMULATED EMISSION OF RADIATION

C. R. DITCHFIELD

Royal Radar Establishment, Malvern

Stimulated emission, produced by interaction between microwave radiation and a molecular system, can result in coherent amplification or oscillation. At these frequencies, spontaneous emission, which causes random electrical noise, is dominated by stimulated emission. Mechanisms which produce excess noise, being absent, the Maser is an extremely low-noise amplifier. The physical processes underlying absorption and radiation, together with the potentialities of gaseous and solid state Masers, are described.

M.A.S.E.R., an acronym for the title of this article, was coined by Professor TOWNES[1] of Columbia University in 1955 to describe the process whereby energy, stored in a molecular or atomic system, could be used to provide coherent amplification or oscillation at microwave frequencies. The subject has two main facets: firstly, the provision of extremely low noise amplifiers and very monochromatic oscillators, which has obvious practical attractions; secondly, the theoretical interest in a process of interaction between radiation and matter which is virtually impossible to observe at optical frequencies, where the physicist has accumulated most of his spectroscopic data. An outline of the theory is presented as an aid to the appreciation of practical Masers and their future possibilities.

Absorption Lines

The quantum theory postulates that an atom (or molecule) can exist only in a series of states corresponding to discrete values of energy. The differences between the energy levels, E_1, E_2, E_3 *etc* decide the frequencies of the radiation which will interact with the atom, according to

$$h\nu_{12} = E_2 - E_1$$
$$h\nu_{23} = E_3 - E_2, etc$$

where h is Planck's constant, $6{\cdot}624 \times 10^{-27}$ erg sec, and ν_{ab} is the frequency appropriate to a transition between levels a and b. If the atom absorbs a quantum, $h\nu$, it jumps from a lower to a higher energy level, the reverse process being accompanied by the emission of a quantum.

For a single atom the energy levels are quite discrete and sharp lines are obtained. However, in an atomic or molecular conglomeration, particularly in a solid, interaction occurs between neighbours, resulting in a broadening of the energy levels and a consequent width to the absorption line. Instead of a single frequency, a narrow band of frequencies can be absorbed (or emitted) and a characteristic line shape is obtained, as in *Figure 1*, when the magnitude of the absorption is plotted against frequency. The details of the shape will depend on the broadening mechanism; the Lorentzian line shape, for example, is represented by

$$g(\nu) = \frac{g(\nu_0)}{1 + \{2\pi(\nu - \nu_0)T_2\}^2}$$

In thermal equilibrium the number of atoms in each level, *i.e.*, the population, is determined by the Boltzmann distribution. Thus for two levels we have a population ratio

$$\frac{N_b}{N_a} = \mathrm{e}^{\frac{-h\nu_{ab}}{kT}}$$

where k is Boltzmann's constant $1{\cdot}38 \times 10^{-16}$ ergs per degree and T is the absolute temperature. For microwave frequencies when $h\nu << kT$, the exponential is often approximated to the first terms of its series.

The probability of a particular atom suffering a transition has the same value for either the absorption or emission of a quantum. Consequently the nett absorption of power for an atomic system is proportional to

$$W_{12}(N_1 - N_2)h\nu_{12}$$

where W_{12} represents the probability of a transition between the levels occurring in unit time. In thermal equilibrium the population N_1 of the lower level is the greater and there is a nett absorption of power. This absorption disturbs the values of the populations, but there is also a tendency for the system to relax to its state of thermal equilibrium by some system of energy interchange. Hence, for small incident powers, the populations do not differ significantly from the equilibrium values. However if the relaxation time is sufficiently long, a large incident power can saturate the transition, that is, make N_1 equal to N_2; no further absorption can take place and the substance is transparent. This process cannot make the population of the upper level exceed that of the lower level, because of the identity of transition probabilities for upward and downward transitions.

The observation of absorption lines in the microwave region has led to the techniques of microwave spectroscopy[2-4], which have been used to elucidate such problems as that of interatomic bonding.

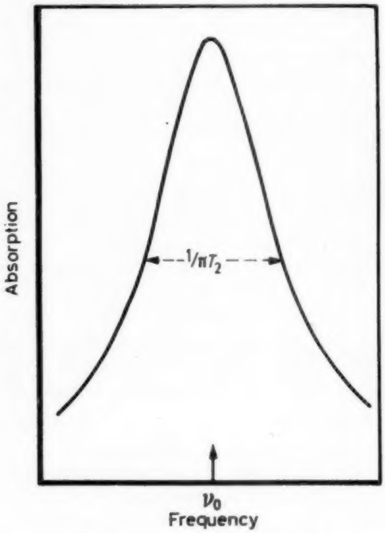


Figure 1. Lorentzian line shape

However, no direct use can be made of this absorptive process to provide amplification of a microwave signal and, although it is basically capable of providing a frequency standard, as discussed by L. Essen[5], the degree of monochromatism is limited by the relatively small population difference between the levels. By making the population of the upper level exceed that of the lower and thus providing stimulated emission of radiation improved amplifiers and oscillators can be obtained.

Stimulated Emission of Radiation

Visible radiation, for example from a neon discharge, is incoherent, being caused by spontaneous emission, a random release of quanta as some of the upper population fall to a lower energy level. Useful amplification demands coherence, whereby each emission is stimulated by the electromagnetic wave. Einstein has shown that the probability ratio of spontaneous to stimulated emission, for thermal equilibrium, is given by

$$e^{-h\nu/kT} - 1$$

This has a large value at optical frequencies, but is small at longer wavelengths; for room temperature and a ten centimetre wavelength, the probability ratio is about 5×10^{-4}. In the microwave region the spontaneous emission, which corresponds to electrical noise, is dominated by the process of stimulated emission, which, in a Maser, can provide either very low noise amplification or an extremely pure spectrum in an oscillator. The problem is to force a population excess into the upper energy level.

The Ammonia Maser

The molecular beam Maser[1], due to C. H. Townes and his colleagues, was the first to be successful. The electric dipole moment of the ammonia molecule results in a pair of energy levels which provide resonant absorption at a wavelength in the region of 1·25 cm. The required population excess is obtained by the action of an inhomogeneous d.c. electric field on a stream of gaseous molecules. A physical separation occurs, according to the energy state of each molecule, so that the higher energy molecules are focussed through a hole in a resonant microwave cavity. The coincidence of the cavity resonance with that corresponding to the energy spacing of the ammonia molecules ensures that they are stimulated to emit radiation, which builds up the microwave field in the cavity. With sufficient molecules present, the molecular power emitted is more than enough to overcome the losses in the cavity and amplification (or oscillation) is possible.

In oscillation the ammonia Maser has provided the most monochromatic source of radiation ever achieved: a spectral purity of a few parts in 10^{12}. The frequency emitted is characteristic of the ammonia molecule and is eminently suitable for a frequency standard, the stability being a considerable improvement on that obtained from an absorption line. Vibration, chemical and thermal changes alter the resonant frequency of the cavity, which in turn, detunes the oscillation frequency, though to a very much smaller extent. Frequency discrimination techniques have been devised to further minimize the effect and a long term stability equivalent to better than a second a century seems likely.

As an amplifier the ammonia Maser has the disadvantages of operating at only one frequency and of providing useful gain over onlya small bandwidth, *i.e.* a few kilocycles per second, with small output power. Some restriction arises inevitably, because of the relatively low concentration of molecules, thus limiting the energy available for signal amplification. Solids provide a greater concentration of active centres and, in particular, use has been made of paramagnetic resonance, wherein the microwave field interchanges energy with the magnetic dipole moment of the electron system. Although this moment has only about one per cent of the value of the electric dipole moment used in the ammonia Maser, the higher concentration more than offsets this and solid state Maser amplifiers have provided bandwidths of tens of megacycles per second. Moreover the resonant frequency of the system can be changed by the application of an external magnetic field, thus providing a tunable amplifier.

Paramagnetic Resonance

The hypothesis of the spinning electron, introduced to explain the multiplicity of spectra such as the sodium doublet, together with the Pauli exclusion principle, can be used to explain the arrangements of the orbital electrons round the nucleus of an element, leading to the Periodic Table. Associated with its spin, s, an electron has a magnetic moment μ. The magnetic moment of an ion is the vector sum of those of its individual electrons. Our particular concern is with the paramagnetic ions of the transition groups, titanium to nickel, and the rare earths. These possess large magnetic moments because of the non-zero spin vector of an incomplete intermediate shell of electrons. (Paramagnetism can also be obtained from excess electrons associated with crystalline lattice defects: those caused by neutron bombardment or by impurities, *e.g.*, phosphorus in silicon.)

In ions such as trivalent chromium, the total orbital angular momentum vanishes and the magnetic moment is that of electron spins, the possible values for chromium being $\pm\frac{3}{2}$ and $\pm\frac{1}{2}$ Bohr magnetons. The chromium ion experiences an internal field due to its neighbours in the crystalline lattice. This causes the $\frac{3}{2}$ values to have an energy different to the $\frac{1}{2}$ values, the extent of this crystalline field splitting depending on the specific lattice. When subjected to an external magnetic field the spin energy depends on the sign of the spin, *i.e.*, whether its moment has a component parallel or antiparallel to the field. This Zeeman splitting thus produces four energy levels for chromium. A simple linear relation holds with the d.c. field along the crystalline axis (*Figure 2a*), but the picture changes as the angle between these two directions is varied (*Figures 2b* to *2d*).

Paramagnetic absorption may occur when the frequency of the incident radiation corresponds to

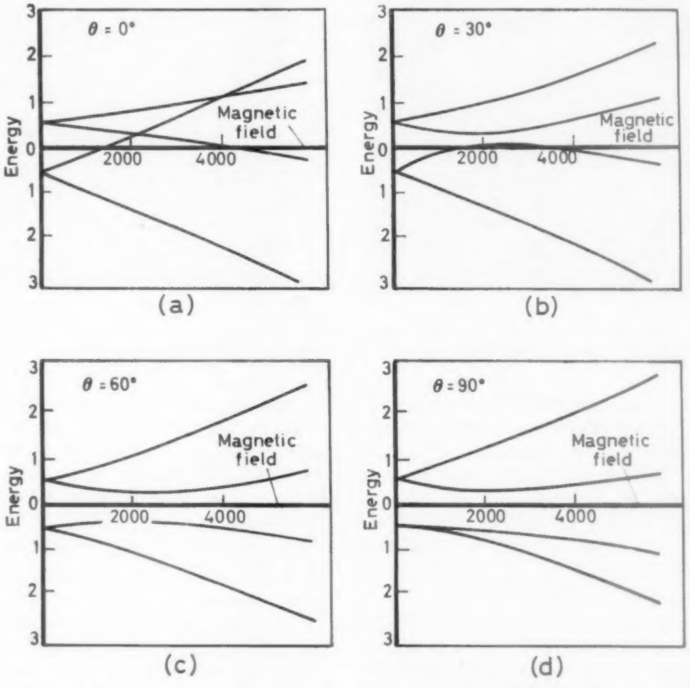


Figure 2. Energy levels of Cr+++ in ruby for various angles θ between the external magnetic field and the c axis of the crystal. Magnetic field in oersteds; energy in kilomegacycles per second to facilitate conversion to transition frequency ($E = h\nu$)

The observation of absorption lines in the microwave region has led to the techniques of microwave spectroscopy[2-4], which have been used to elucidate such problems as that of interatomic bonding.

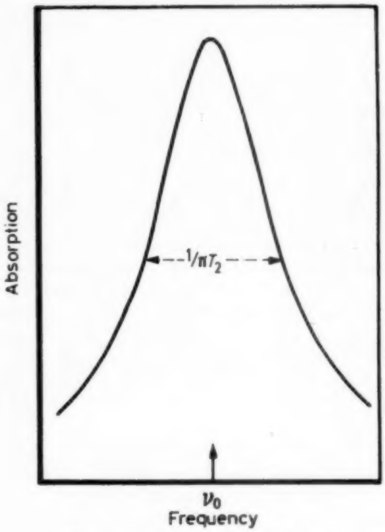


Figure 1. Lorentzian line shape

However, no direct use can be made of this absorptive process to provide amplification of a microwave signal and, although it is basically capable of providing a frequency standard, as discussed by L. ESSEN[5], the degree of monochromatism is limited by the relatively small population difference between the levels. By making the population of the upper level exceed that of the lower and thus providing stimulated emission of radiation improved amplifiers and oscillators can be obtained.

Stimulated Emission of Radiation

Visible radiation, for example from a neon discharge, is incoherent, being caused by spontaneous emission, a random release of quanta as some of the upper population fall to a lower energy level. Useful amplification demands coherence, whereby each emission is stimulated by the electromagnetic wave. Einstein has shown that the probability ratio of spontaneous to stimulated emission, for thermal equilibrium, is given by

$$e^{-h\nu/kT} - 1$$

This has a large value at optical frequencies, but is small at longer wavelengths; for room temperature and a ten centimetre wavelength, the probability ratio is about 5×10^{-4}. In the microwave region the spontaneous emission, which corresponds to electrical noise, is dominated by the process of stimulated emission, which, in a Maser, can provide either very low noise amplification or an extremely pure spectrum in an oscillator. The problem is to force a population excess into the upper energy level.

The Ammonia Maser

The molecular beam Maser[1], due to C. H. TOWNES and his colleagues, was the first to be successful. The electric dipole moment of the ammonia molecule results in a pair of energy levels which provide resonant absorption at a wavelength in the region of 1·25 cm. The required population excess is obtained by the action of an inhomogeneous d.c. electric field on a stream of gaseous molecules. A physical separation occurs, according to the energy state of each molecule, so that the higher energy molecules are focussed through a hole in a resonant microwave cavity. The coincidence of the cavity resonance with that corresponding to the energy spacing of the ammonia molecules ensures that they are stimulated to emit radiation, which builds up the microwave field in the cavity. With sufficient molecules present, the molecular power emitted is more than enough to overcome the losses in the cavity and amplification (or oscillation) is possible.

In oscillation the ammonia Maser has provided the most monochromatic source of radiation ever achieved: a spectral purity of a few parts in 10^{12}. The frequency emitted is characteristic of the ammonia molecule and is eminently suitable for a frequency standard, the stability being a considerable improvement on that obtained from an absorption line. Vibration, chemical and thermal changes alter the resonant frequency of the cavity, which in turn, detunes the oscillation frequency, though to a very much smaller extent. Frequency discrimination techniques have been devised to further minimize the effect and a long term stability equivalent to better than a second a century seems likely.

As an amplifier the ammonia Maser has the disadvantages of operating at only one frequency and of providing useful gain over onlya small bandwidth, *i.e.* a few kilocycles per second, with small output power. Some restriction arises inevitably, because of the relatively low concentration of molecules, thus limiting the energy available for signal amplification. Solids provide a greater concentration of active centres and, in particular, use has been made of paramagnetic resonance, wherein the microwave field interchanges energy with the magnetic dipole moment of the electron system. Although this moment has only about one per cent of the value of the electric dipole moment used in the ammonia Maser, the higher concentration more than offsets this and solid state Maser amplifiers have provided bandwidths of tens of megacycles per second. Moreover the resonant frequency of the system can be changed by the application of an external magnetic field, thus providing a tunable amplifier.

Paramagnetic Resonance

The hypothesis of the spinning electron, introduced to explain the multiplicity of spectra such as the sodium doublet, together with the Pauli exclusion principle, can be used to explain the arrangements of the orbital electrons round the nucleus of an element, leading to the Periodic Table. Associated with its spin, s, an electron has a magnetic moment μ. The magnetic moment of an ion is the vector sum of those of its individual electrons. Our particular concern is with the paramagnetic ions of the transition groups, titanium to nickel, and the rare earths. These possess large magnetic moments because of the non-zero spin vector of an incomplete intermediate shell of electrons. (Paramagnetism can also be obtained from excess electrons associated with crystalline lattice defects: those caused by neutron bombardment or by impurities, *e.g.*, phosphorus in silicon.)

In ions such as trivalent chromium, the total orbital angular momentum vanishes and the magnetic moment is that of electron spins, the possible values for chromium being $\pm \frac{3}{2}$ and $\pm \frac{1}{2}$ Bohr magnetons. The chromium ion experiences an internal field due to its neighbours in the crystalline lattice. This causes the $\frac{3}{2}$ values to have an energy different to the $\frac{1}{2}$ values, the extent of this crystalline field splitting depending on the specific lattice. When subjected to an external magnetic field the spin energy depends on the sign of the spin, *i.e.*, whether its moment has a component parallel or antiparallel to the field. This Zeeman splitting thus produces four energy levels for chromium. A simple linear relation holds with the d.c. field along the crystalline axis (*Figure 2a*), but the picture changes as the angle between these two directions is varied (*Figures 2b* to *2d*).

Paramagnetic absorption may occur when the frequency of the incident radiation corresponds to

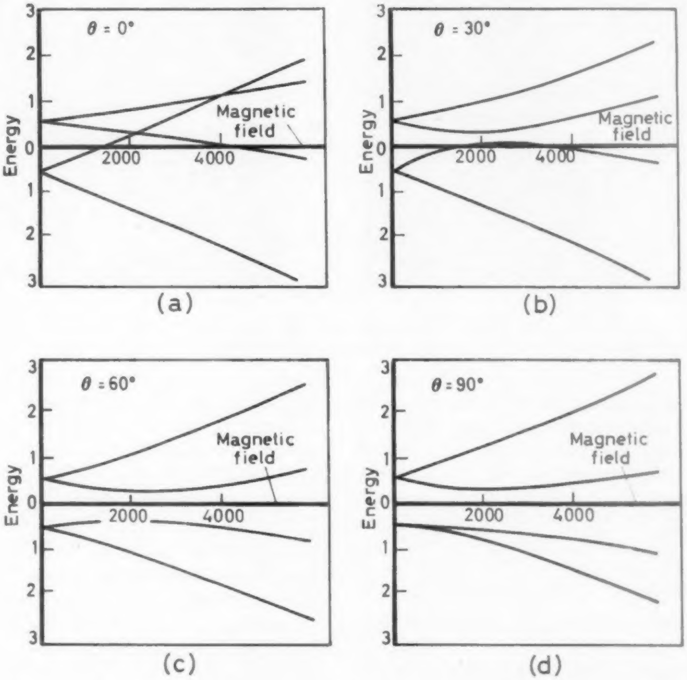


Figure 2. Energy levels of Cr^{+++} in ruby for various angles θ between the external magnetic field and the c axis of the crystal. Magnetic field in oersteds; energy in kilomegacycles per second to facilitate conversion to transition frequency ($E = h\nu$)

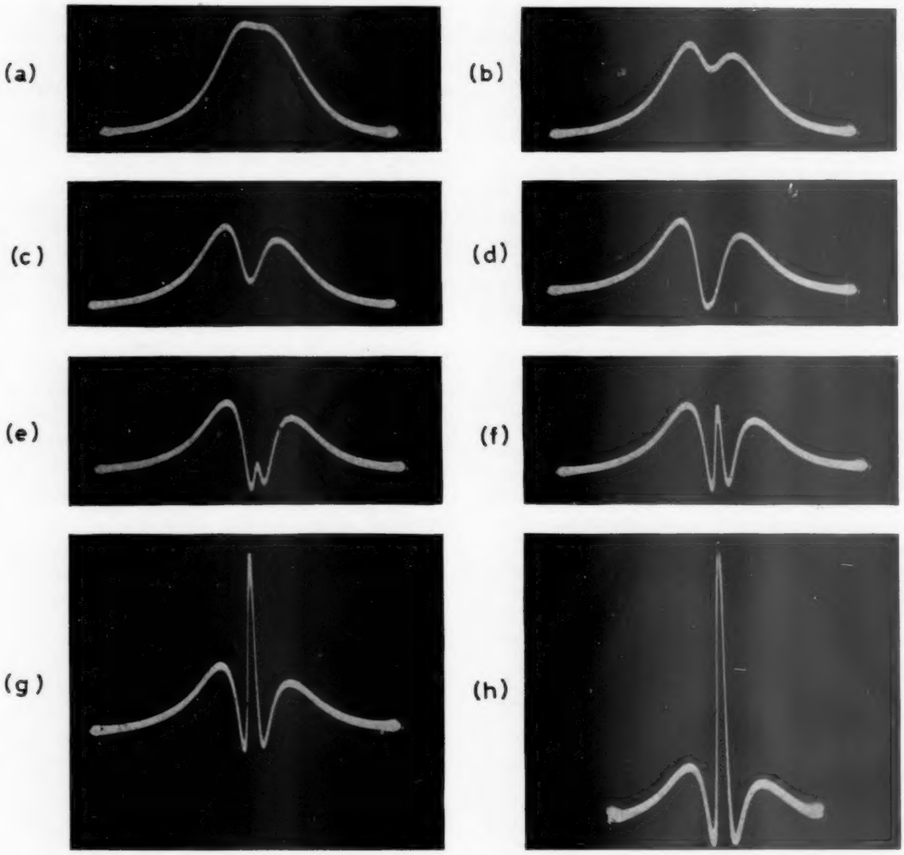


Figure 3. The effect on the signal absorption line of successively higher levels of pump power. These photographs were obtained using the R.R.E. ruby Maser working with liquid oxygen as the refrigerant

the energy gap between two levels; pictorially when this frequency coincides with the Larmor precession frequency of the electron gyromagnet. Where the energy levels are non-linearly related to the magnetic field, they cannot be characterized by a single quantum number (an 'admixture of states') and transitions can occur which would be forbidden, normally, by the selection rules. This fact is fundamental to the concept of the three level Maser.

The proximity of a similar neighbour influences a paramagnetic ion, the effect being to broaden the absorption line as the ions are brought closer together. For Maser purposes the active ion is 'diluted' by an isomorphous diamagnetic ion, *e.g.*, in ruby, an aluminium oxide lattice with one aluminium ion in, say, a thousand replaced by paramagnetic chromium. The degree of dilution is a compromise between the advantage of having the maximum number of active centres per unit volume and the disadvantage of having absorption lines so broad that the energy levels cannot be treated as ideal, discrete and independent states.

The Two Level Solid State Maser

In a system with only two levels, *i.e.*, spin $\pm \frac{1}{2}$, most of the spins, N_1, will be in the lower energy state with their magnetic moment μ parallel to the field and fewer spins N_2 will have their moments antiparallel. Various methods of achieving a population excess in the upper level can be postulated, the simplest to visualize being a reversal of the d.c. field in a time short compared with the relaxation time. This experiment was performed by E. M. PURCELL and R. V. POUND using nuclear spins, but these have moments about a thousand times smaller than those of electrons and the stimulated emission available did not exceed the circuit losses.

An easier process to operate is that of adiabatic fast passage, wherein the frequency of an oscillator is swept rapidly from one side of the resonance value to the other. This inverts the individual moments and thus has essentially the effect of reversing the population values in the upper and lower states. The atomic system is now able to provide stimulated emission of radiation, but the gain will decrease as the relaxation process tends to restore the populations to the values corresponding to thermal equilibrium. Cooling the system has two advantages: it increases the relaxation time and it decreases the limiting noise temperature of the amplifier. The first attempt[6] to obtain Maser action in the solid state used the two levels present in phosphorus doped silicon cooled to 2°K. The two level Maser provides only intermittent operation and much more effort has gone into the three level Maser.

The Three Level Solid State Maser

In the three level Maser, as first propounded by Professor BLOEMBERGEN[7] in 1956, power is applied at a frequency corresponding to the transition between non-adjacent energy levels, for example, 1 to 3. There is a consequential tendency to equalize the populations in these two levels, with appropriate changes in N_3/N_2 and N_2/N_1, as the pump power between the outer levels is increased. Saturation of the pump transition, that is $N_1 = N_3$, ensures that either $N_3 > N_2$ or $N_2 > N_1$. The alternative occurring in practice, for any particular substance, depends on the relative position of the intermediate level and on the transition probabilities between the levels. These probabilities depend on the strength and orientation of the microwave magnetic field and also on the transition involved.

The effect on the signal absorption line (in this instance between levels 2 and 3) of successively higher pump powers is shown in *Figure 3*. It can be seen that the absorption in the centre of the line is decreased until the substance is transparent: $N_2 = N_3$ and then, as the ratio N_3/N_2 increases, more stimulated emission becomes available, the circuit losses are overcome and the system amplifies.

The process provides continuous operation and the noise temperature of the signal transition, decided by the ratio N_3/N_2, can be a fraction of ambient temperature. Advantage is again obtained by cooling the system to decrease noise and to obtain long relaxation times, thus decreasing the pump power required, only a few microwatts being necessary for ruby at helium temperatures.

Although the fundamental idea of the three level Maser has the beauty of being essentially simple, some of the finer details have a complexity which has not been elucidated. Nevertheless many three level Masers are now working. The main centre of effort is the United States, with smaller numbers in the United Kingdom, Australia and Holland. Some of these achievements will serve to illustrate practical considerations of Maser design and to indicate the performance to be expected.

Practical Three Level Maser Systems

In the first successful three level Masers[8-13] the interaction between the microwave field and the spin system was increased by placing the paramagnetic sample in a cavity, which was resonant simultaneously to the pump and signal frequencies. An external field of appropriate magnitude and orientation ensured that the transitions in the crystal corresponded to these frequencies. Amplification is possible when the power available from stimulated emission exceeds that absorbed in the cavity walls; the device oscillates if the stimulated emission is sufficient to overcome also the power lost through the coupling between the cavity and the remainder of the microwave system. Expressed in terms of the Q factor, defined as

$$Q = 2\pi\nu \cdot \frac{\text{stored energy}}{\text{power absorbed}}$$

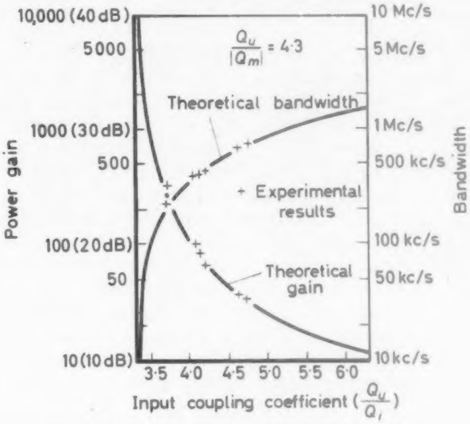


Figure 4. Maser gain and bandwidth as a function of input coupling. The experimental points were obtained with the first Maser working near to 60°K. Better gain-bandwidths would be obtained at lower temperatures

the expression for stable amplification may be written as

$$\frac{1}{Q_u}+\frac{1}{Q_1}>\frac{1}{|Q_m|}>\frac{1}{Q_u}$$

where Q_u is the unloaded Q factor of the cavity, Q_1 is the Q factor of the coupling aperture and Q_m is the magnetic Q of the sample. Q_m is negative in sign for an emissive sample.

To achieve good performance $|Q_m|$ should be as low as possible, as this determines the bandwidth and gain which can be obtained. For a reflection cavity Maser, in which the incident signal is amplified and reflected back along the input line in the reverse direction, the formulae for bandwidth B, and power gain G, are

$$B=\left[\frac{1}{Q_1}-\left(\frac{1}{|Q_m|}-\frac{1}{Q_u}\right)\right]v$$

$$G=\left[\frac{\frac{1}{Q_1}+\left(\frac{1}{|Q_m|}-\frac{1}{Q_u}\right)}{\frac{1}{Q_1}-\left(\frac{1}{|Q_m|}-\frac{1}{Q_u}\right)}\right]^2.$$

Therefore the quantity

$$B\left[\sqrt{G}-1\right]=2v\left[\frac{1}{|Q_m|}-\frac{1}{Q_u}\right]$$

is a constant for a particular system. Thus the bandwidth may be increased, with a consequent decrease in gain, by increasing the loading on the cavity, that is decreasing Q_1. This is shown graphically in *Figure 4*. The system described has the disadvantage that it requires a circulator to distinguish between the incident and amplified signals. A transmission cavity Maser, for which essentially similar expressions can be derived, requires only an isolator at its output to prevent noise from subsequent components in the system reaching the Maser and being amplified. However its gain–bandwidth product for the same value of $|Q_m|$ is somewhat worse than the reflection cavity Maser, which has been preferred in experiments to date.

Only a few months after the publication of Bloembergen's proposal for a three level Maser, workers at the Bell Telephone Laboratories, announced[8] oscillations at 9·06 kMc/s using a pump power of 17·52 kMc/s. The salt used was gadolinium ethyl sulphate at 0·5 per cent dilution in a lattice of lanthanum ethyl sulphate, and was contained in a resonant cavity immersed in liquid helium at 1·2°K, with a magnetic field of 2850 oersteds. A rectangular cavity was used resonant in an H_{01} mode at 17·5 kMc/s, and a strip line mode at 9·06 kMc/s with the salt placed as shown in *Figure 5*.

On a simple theory of the electron spin gyro, a transition between adjacent levels has the maximum probability when the r.f. field is perpendicular to the d.c. field, as it can couple most effectively with the component of spin moment precessing at the Larmor frequency. In general, with the admixture of quantum states, the transition probabilities for signal and pump frequencies are a maximum at particular orientations of the respective r.f. fields with relation to the d.c. field. These optima are calculable from the matrix elements of the wave functions corresponding to the states of the paramagnetic ion. A rectangular cavity has the advantage that the direction of the r.f. field can be substantially the same over quite a large sample. Therefore spins in all parts of the crystal will react most efficiently with the pump and signal r.f. fields, particularly if the transition probability requirements are for parallel or orthogonal fields, which can be conveniently arranged in such a cavity.

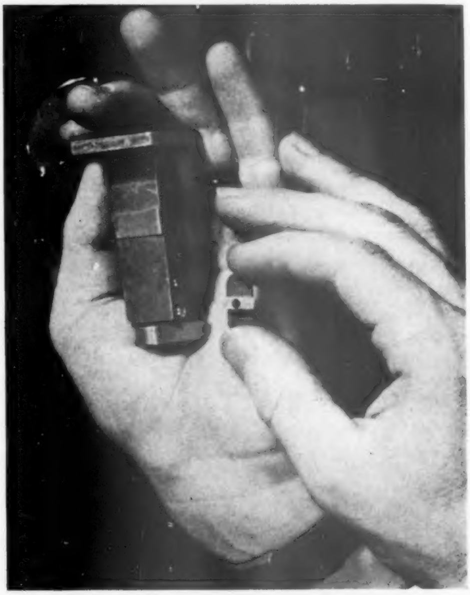


Figure 5. A Maser cavity (by courtesy of The Editor of Bell Laboratories Record)

In the Bell Telephone Laboratories cavity the r.f. fields would have been optimum when orthogonal, whereas they were, in fact, parallel. Consequently the d.c. field was arranged to be at 45° to the r.f. fields, so that the signal field had a component perpendicular and the pump field a component parallel to the external magnetic field. In a cavity of circular cross section the direction of the field will change over the sample and only a component of it is really effective. Thus, two factors which affect $|Q_m|$ are the orientation of the pump field, which affects the extent to which this transition may be saturated and the orientation of the signal field which determines the ease with which stimulated emission is obtained from the spins.

A filling factor is defined as the proportion of the energy in the cavity which would be stored in the volume occupied by the sample, that is

$$\eta = \frac{\int H_s^2 \mathrm{d}v}{\int H_s^2 \mathrm{d}V}$$

where v and V refer to the sample and the cavity volumes. The pump and signal fields will vary in intensity throughout the cavity to a different extent and it would appear that the advantage of a large filling factor is offset if part of the crystal contains a node at the pump frequency, as, in this region, the populations would not be disturbed. In practice completely filled cavities have been used, a possible explanation being that phonons can be responsible for energy interchange between different parts of the crystal, thus resulting in an adequate population ratio over the whole volume.

The final contributions to $|Q_m|$ arise from the available spin population per unit volume per unit line width, effectively NT_2, and their distribution to give a large ratio between the populations in the states apertaining to the signal transition. Because the state populations decrease exponentially with the energy difference, *i.e.*, in proportion to the frequency of transition, there is advantage in using a high value of pump frequency, as, other things being equal, this will produce a large ratio of the signal level populations.

A very interesting variation on this is the method of 'push–pull pumping', which can be applied particularly when two suitable pump frequency transitions occur at the same frequency, for a given magnetic field. For example, at a particular angle of 54° 44′ between the d.c. field and the crystalline axis of ruby, the energy levels are symmetrical, as in *Figure 6*. Thus pump power applied between levels 1 and 3 will also act between levels 2 and 4, pushing spins into level 3 and pulling them out of level 2. The result is a value of $|Q_m|$ which is several times lower than can be obtained by simple pumping between one pair of levels. With other ions such as Fe^{3+} in sapphire, degeneracy can occur between more than two transitions, thus giving further scope to this idea.

So far it has been assumed that the width of the absorption line, expressed as a frequency, is much greater than the amplifying bandwidth but with more recent developments in practical Maser systems this is not so. Therefore, in addition to $|Q_m|$ a factor Q_s apertaining to the absorption line and defined as πT_{2v} must be taken into account. If $|Q_m|$ is very low, Q_s is the limiting factor in bandwidth performance and it could be advantageous to broaden the absorption line, that is to decrease T_2, by use of an inhomogeneous d.c. field, thus adjusting Q_s to be approximately the same value as $|Q_m|$.

Cavity Masers have been constructed which are tunable[10], or can operate at relatively high temperatures[13], or can provide gain bandwidth products[10] of nearly 100 Mc/s, that means, a bandwidth of several megacycles per second with a gain of 20 dB. Further research, particularly into paramagnetic materials, might lead to all these features being combined in one device, with perhaps only a very small external magnet, particularly if a large splitting can be obtained from the crystalline field.

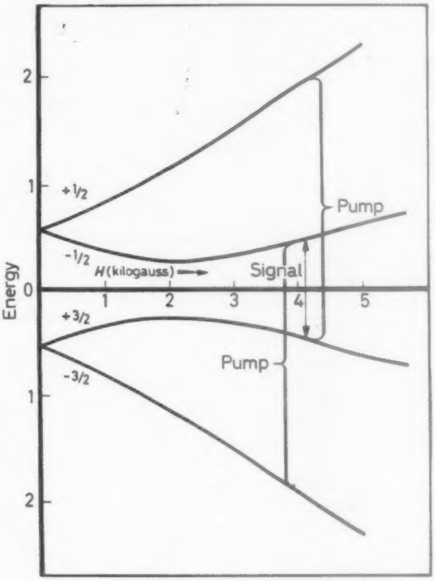


Figure 6. Energy levels of Cr+++ in ruby. Magnetic field at 54° 44′ to c-axis. (Energy expressed in kMc/s)

However, another type of microwave circuit, the travelling wave structure, has many advantages over a cavity *e.g.* it has good gain stability, tunability, its amplifying bandwidth is less limited by Q_s and the unidirectional property, provided by an external device in a cavity system, can be built into the Maser itself. Such a Maser has been built[14] by Bell Telephone Laboratories (*Figure 7*). The wave velocity is decreased by a comb-like structure which effectively increases the interaction between the wave and the crystals. The circular polarization of the wave, which has opposite sense either side of the comb, is cleverly utilized to give amplification in one direction, when the wave interacts with the Larmor precession of ruby with a low chromium concentration, but attenuation for a wave in the other direction, which interacts with heavily doped ruby, in which the populations are not brought into the emissive condition by the pump power. A gain of about 25 dB over a bandwidth of 25 Mc/s has been obtained.

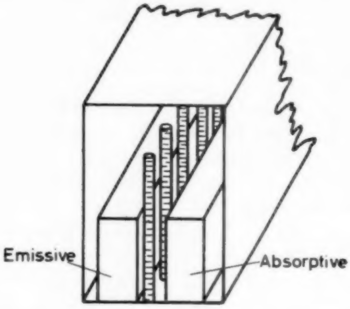


Figure 7. Diagram of travelling wave structure in a ruby Maser used at Bell Telephone Laboratories

Noise in Masers

It has been customary to judge receivers by their noise factor. This is an adequate criterion for relatively poor receivers, *e.g.*, of noise factor six times or more, where the output noise is little affected by the background temperature against which a signal is to be detected. However, with low noise amplifiers, such as Masers, the reverse is true and a much better criterion is the effective noise temperature contribution of the Maser amplifier. It can be shown that

$$T_{eff}=\frac{4}{Q_1|Q_m|\left[\frac{1}{Q_1}+\frac{1}{|Q_m|}-\frac{1}{Q_u}\right]^2}\cdot\left[|T_m|+\frac{|Q_m|}{Q_u}T_c\right]$$

for a reflection cavity, where T_c is the temperature of the cavity and T_m is defined in terms of the emissive populations of the signal transition energy states. A large ratio decreases $|T_m|$ as well as $|Q_m|$, thus emphasizing again the advantage of such schemes as push–pull pumping. These can result in $|T_m|$ being below ambient temperature. The sensitivity limit of a three level Maser can thus be below the physical temperature of the device.

In practice the noise performance of a Maser is so good that it is extremely difficult to measure accurately, being about a thousand times lower than that of a conventional receiver. However an upper limit of 10° to 20°K has been set by A. L. McWhorter, J. W. Meyer and P. D. Strum[15] of Lincoln Laboratories on a cavity Maser using potassium chromicyanide and by R. W. DeGrasse, E. O. Schulz-DuBois and H. E. D. Scovil, of Bell Telephone Laboratories[14], on the travelling wave ruby Maser.

Thus there is little doubt that a system using a Maser will be limited by noise contributions from other parts of the equipment and work will be necessary to improve aerials and transmission circuits from this point of view.

Application of Masers

To date Masers have been developed mainly for their own interesting properties, with obvious application in the future to low noise amplifiers used in scientific investigations such as microwave spectrometry and radio astronomy. In the latter field the Naval Research Laboratories, Washington, in cooperation with Columbia University, have used a Maser as a pre-amplifier for a radio telescope, with great advantage over a conventional receiver. Other applications lie in the realms of radar and long distance communication, perhaps the most exciting being those techniques concerned with global communications via satellites and, eventually, inter-planetary travel. Finally, although all present systems involve use of a pump frequency higher than that at which stimulated emission is obtained, several possibilities exist whereby coherent oscillators and amplifiers may be obtained at submillimetre wavelengths, which region is at present undeveloped.

The extent to which Masers are used will depend on the extent to which their low noise properties can be utilized effectively in a system and on the degree of complexity in the Maser itself. At present the provision of useful gain over large bandwidths entails very low temperatures but, in a multi-level Maser, this in not fundamentally necessary. A large population ratio for an emissive signal transition can be obtained by pumping several pairs of levels and by a favourable adjustment of the relaxation times between specific energy states. The latter technique would also improve the power handling capacity of a Maser, which, although completely adequate for passive receivers, has so far limited application to radar. In gadolinium ethyl sulphate one transition time was changed by an order of magnitude[61] on introducing a small number of cerium ions into the lattice.

Although data exist on the paramagnetic properties of many materials, the precise mechanism of the spin–spin and spin–lattice relaxation processes needs further elucidation. This could lead to the synthesis of a material with desired properties, which, 'tailored' for Maser applications, might transcend present crystals to such an extent that low noise, adequate gain and wide bandwidth will be obtainable from a relatively simple Maser working at easily achievable temperatures, such as 77°K (liquid nitrogen), with only a very small external magnetic field for purposes of tuning.

In conclusion it is apparent that Maser techniques have developed rapidly in the few years since their conception and it would appear that the ultimate possibilities still remain to be exploited and dividends should continue to accrue from further research.

This article is published by kind permission of the Director of the Royal Radar Establishment and the Ministry of Supply.

References

[1] GORDON, J. P., ZEIGER, H. J. and TOWNES, C. H. *Phys. Rev.* (1955) **99,** 1264

[2] GORDY, W., SMITH, W. V. and TRAMBARULO, R. F. *Microwave Spectroscopy* London: Chapman and Hall, 1953

[3] TOWNES, C. H. and SCHAWLOW, A. L. *Microwave Spectroscopy* New York: McGraw Hill, 1955

[4] INGRAM, D. J. E. *Spectroscopy at Radio and Microwave Frequencies* London: Butterworths Scientific Publications, 1955

[5] ESSEN, L. *Research, Lond.* (1957) **10,** 217

[6] COMBRISSON, J., HONIG, A. and TOWNES, C. H. *C. R. Acad. Sci., Paris* (1956) **242,** 2451

[7] BLOEMBERGEN, N. *Phys. Rev.* (1956) **104,** 324

[8] SCOVIL, H. E. D., FEHER, G. and SEIDEL, H. *ibid.* (1957) **105,** 762

[9] MCWHORTER, A. L. and MEYER, J. W. *ibid.* (1958) **109,** 312

[10] STRANDBERG, M. W. P. *et al. Proc. Inst. radio Engrs* (1959) **47,** 1, 80

[11] ARTMAN, J. O., BLOEMBERGEN, N. and SHAPIRO, S. *Phys. Rev.* (1958) **109,** 1392

[12] MAKHOV, G., KIKUCHI, C., LAMBE, J. and TERHUNE, R. W. *ibid.* p. 1399

[13] DITCHFIELD, C. R. and FORRESTER, P. A. *Phys. Rev. (Letters)* (1958) **1,** 448

[14] DEGRASSE, R. W., SCHULZ-DUBOIS, E. O. and SCOVIL, H. E. D. *Bell System Tech. J.* (1959) **38,** 305

[15] MCWHORTER, A. L., MEYER, J. W. and STRUM, P. D. *Phys Rev.* (1957) **108,** 1642

[16] FEHER, G. and SCOVIL, H. E. D. *ibid.* (1957) **105,** 760

Book Reviews

Preparation of Single Crystals

W. D. LAWSON and S. NIELSON

(vii+255 pp; 8¾ in. by 5½ in.)

London: Butterworths Scientific Publications: New York: Academic. 45*s*; $8.80

THIS book is one of a projected series of semiconductor monographs and is concerned with the growth of semiconducting crystals. The general recognition of the importance of crystal purity and perfection in the experimental study and application of these solids, and the consequent interest in the improvement of techniques of preparation, make this a subject of great current interest.

An initial elementary outline of some electrical properties of these materials is followed by a survey of methods of crystal growth and a description of equipment. The authors then discuss means of purification and analysis, and procedures for the preparation of compounds.

An account of equilibrium and non-equilibrium defects leads to a discussion of the mechanisms of growth, together with some further techniques. A final chapter, with a title open to misconstruction, is mainly concerned with the interaction of impurities with the lattice.

The book has an air of informality, makes very few demands on the reader in the form of knowledge of semiconducting systems, and provides a clear and valuable account of the techniques. The authors sometimes go to great lengths to ensure that practical and constructional details are given. Many of these are very valuable but it does not seem possible to justify the inclusion of some of the material. For example, in the chapter on apparatus, the elementary theory of the McLeod gauge and diagrams illustrating the principles of operation of rotary backing pumps are out of place.

More generally, the selection and emphasis of the contents may in some way restrict the value of the book. The primary aim is clearly to provide a practical account of techniques, particularly for the newcomer to the subject. This aim is very well satisfied. The authors do subsequently give a simple but satisfactory discussion of growth mechanisms, but do not analyse their objectives. Nor do they discuss the design of procedures to produce given physical properties. In the field of semiconductors the conflicts between different practical objectives or of criteria of perfection, would surely have been of interest to the reader, however practical his approach, and one is sorry that these have not been discussed. P. C. BANBURY

See also page 286

THE THERMODYNAMIC BEHAVIOUR OF SOLIDS

M. A. JASWON

Department of Mathematics, Imperial College of Science and Technology, London

In addition to the four articles on crystal physics by Dr JASWON published* during 1958, a fifth has now been added. This provides an outline of the main principles of thermodynamics, particularly in relation to the light they throw on the behaviour of solids. (*The five articles on crystal physics will shortly be reprinted in the form of a paper-covered booklet.*)

A THERMODYNAMIC system can exist in several distinct kinds of state relative to its environment, but they fall into two main classes, equilibrium states and non-equilibrium states. The first possibility means that the system has no tendency to change with time in any way, though to check this rigorously would require an exhaustive series of experiments. For instance, a metal rod might appear to be in equilibrium at a particular temperature and pressure, as judged by the fact that it persisted indefinitely as such. However, if it had previously been cold-worked, or quenched, or subjected to radiation damage, its microcrystalline structure would certainly be changing for some time afterwards. The change of internal structure can sometimes be followed by x-ray analysis, as shown by *Figure 11* (page 19) referring to the self-recovery at room temperature of cold-worked silver filings. This process of recovery essentially involves the elimination of inhomogeneous microstress fields, and therefore the introduction of increasing regularity throughout the lattice, until a limiting degree of order has been achieved. Cold-worked metal thus has the status of a non-equilibrium system relaxing towards equilibrium at a slow but steady rate, over an appreciable time interval. Similar considerations apply to the elimination of quenched-in lattice defects or quenched-in disorder and, in fact, to any process involving the nucleation of an equilibrium system at low temperatures. Generally speaking, the reverse process of high temperature nucleation takes place much more rapidly. For instance a block of ice suddenly heated beyond 0°C will immediately transform into water, so that the non-equilibrium phase —ice above the melting point—has only a fleeting existence.

Perhaps the most important practical example of a non-equilibrium system is provided by glass, which possesses a non-crystalline structure of somewhat ill defined character. According to the recent geometrical considerations of J. D. BERNAL[1], a large number of equal spheres can be close packed in two—and only two—fundamentally different ways. These conform to long range and short range order respectively. No intermediate degree of order is admissible. If so, the existence of two condensed phases of matter becomes understandable and we may also infer that the atomic structure of glass must be closely allied to the structure of liquids. Yet, paradoxically, many properties of glass are much more reminiscent of crystalline solids than of liquids. For instance, Poisson's ratio has the values 0·245, 0·270 and 0·5 for glass, steel and liquids respectively. A more direct indication is given by the specific heat curve, sketched in *Figure 1* for glycerol. As the temperature of the melt, or supercooled liquid, is reduced through a small interval about the transition temperature T_g, the specific heat sharply declines to a value not very different from that of the crystalline phase.

This behaviour at first sight admits to two interpretations, of almost opposite significance. It could indicate an equilibrium order–disorder transition of some internal degree of freedom, *e.g.*, molecular rotation, analogous to the various lambda transitions in solids. This possibility has been considered unlikely by Sir FRANCIS SIMON and his school[2]. In any case such a view implies the almost untenable idea of the glassy phase as a possible equilibrium phase of condensed matter that competes with the crystalline phase down to the lowest temperatures. On the second interpretation, now generally accepted, the fall in specific heat indicates the cooperative freezing-in of whatever order has been attained by the system at T_g. Attainment of further order, in the direction of increasing crystallinity, is almost completely inhibited below T_g, so that the system persists in its non-equilibrium state. More technically expressed, the configurational entropy of the system is not progressively eliminated on lowering the temperature below T_g, but remains locked up in the system even at 0°K. This behaviour stands in marked contrast to that of equilibrium systems

* *Research, Lond.* (1958) **11**, 26, 108, 227, 315

where, by virtue of the Nernst heat theorem (third law of thermodynamics) the entropy necessarily vanishes at 0°K.

A two dimensional model of a crystal is drawn in *Figure 2*. This depicts the motif unit as a pair of atoms linked by a strong, localized bond of covalent or ionic character. Neighbouring units, whose position and orientation must conform to the translational periodicity of the crystal lattice, are linked by secondary, weaker bonds of the van der Waals type. The main difference between a crystal and a glass springs from the fact that neighbouring units in the latter do not conform to any lattice symmetry, but remain fixed at substantially the configuration appropriate to a supercooled liquid at temperature T_g. This deviation from regularity increases the potential energy of the structure, but the increase is energetically tolerable since it affects only the secondary bonds. Evidently, however, only molecular units of sufficient complexity, capable of interacting by a wide variety of weak coupling mechanisms, have a chance of forming a disordered array of the requisite stability.

No equation of state, such as that connecting the temperature, pressure and volume of a crystal, exists for a glass. In the first place, the internal structure of the latter, with derivative physical properties, is not uniquely defined by the temperature and pressure, but depends on the whole previous thermal and mechanical history of the specimen. In the second place, relaxation effects can never be entirely absent in a non-equilibrium system, with the result that its physical properties alter over long intervals of time. On heating a glass, the thermal energy of the molecules soon becomes comparable with the interaction energy of the secondary bonding forces, so that they acquire greater potentialities for relative displacement. Eventually the molecules become sufficiently mobile for the substance to behave identifiably as a liquid, though the precise point at which it ceases to be a solid cannot be sharply determined. It may be remarked that no generally accepted, quantitative theory either of glass or liquids has yet been formulated.

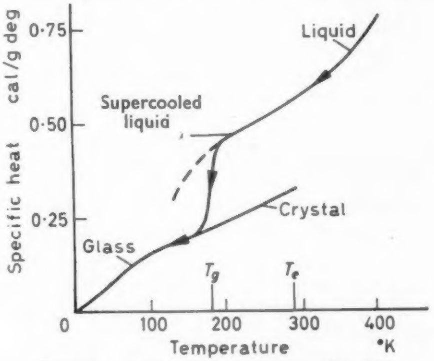


Figure 1. Temperature variation of specific heat of glycerol (after Simon and Lange, cf. Reference 2) indicating freezing temperature T_e and glass transition temperature T_g—the dashed lines indicate expected specific heat if glassy phase did not intervene at T_g

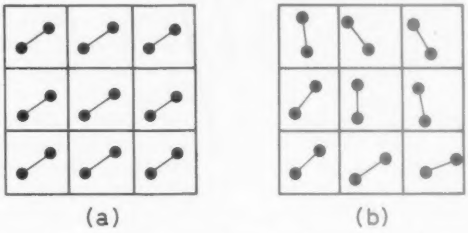


Figure 2. Two dimensional model of (a) crystal and (b) glass

The Entropy Principle

When a system undergoes physico-chemical change, its internal energy usually also changes. Part of this change is balanced by a flow of heat between the system and its surroundings; the remaining part is balanced by mechanical interaction. This is the first law of thermodynamics. In the usual symbolism,

$$dU = dQ + dW, \qquad \dots (1)$$

where all the energy increments are formally positive and acquired by the system. If the volume of the system changes by dV under an external pressure P, then dW equals $(-PdV)$.

Considerations of energy balance do not by themselves suffice to establish the direction of physico-chemical change. This role in thermodynamics is assigned to the entropy function S. A direct evaluation of this function lies beyond the scope of classical thermodynamics, which is limited to the theorem that entropy flow dQ/T accompanies a heat flow dQ. Let us now suppose that S may, by some means, be computed for two neighbouring states of a system: the increment dS then accounts for the net change in entropy of the system arising from the various interacting factors—configurational ordering or disordering of the atoms, volume and shape deformations, heat flow—that make up the physico-chemical process in question. Heat flow entropy is certainly included in dS, but does not generally comprise the whole of dS, for entropy may be created or destroyed within the system quite apart from that passing through the boundaries. On the other hand, the corresponding entropy change in the surroundings can be nothing other than $(-\mathrm{d}Q/T)$. We now formulate the second law of thermodynamics as

$$\mathrm{d}S - \mathrm{d}Q/T \geqslant 0, \qquad \ldots.(2)$$

which means that the net change in entropy of system plus surroundings, considered as a single physical unit, must be either positive or zero.

The equality sign in (*2*) defines reversible processes. In effect, entropy flow here accounts for the whole of dS, so that reversing the flow reverses the sign of dS and hence the direction of change. Alternatively expressed, reversible processes merely involve a redistribution of existing entropy between the system and surroundings, though with no preferred flow in any particular direction: the system thus remains in equilibrium with its surroundings at all stages of the operative process.

Reversible changes do not occur in nature but are important theoretically, sometimes enabling dS to be computed for a given change of state. Consider for example the isothermal expansion of a perfect gas from V to $V+\mathrm{d}V$. We suppose heat is absorbed by the gas and entirely converted into work done by the gas in expanding against the external pressure. The first law of thermodynamics thus reads

$$\mathrm{d}U = \mathrm{d}Q + \mathrm{d}\,W = 0,$$

with $\mathrm{d}W = -P\mathrm{d}V = (-RT/V)\,.\,\mathrm{d}V,$

by virtue of Boyle's equation of state. For reversible changes, we have

$$\mathrm{d}S - \mathrm{d}Q/T = 0, \qquad ..(3)$$

whence $\mathrm{d}S = \mathrm{d}Q/T = -\mathrm{d}W/T = (R/V)\,.\,\mathrm{d}V \quad ..(4)$

From (*4*), the total change in entropy corresponding to a finite change in volume at constant temperature T may readily be calculated. This simple example illustrates an important technique of classical thermodynamics: dS is evaluated indirectly by envisaging a reversible process which connects the relevant final and initial states of the system.

The inequality sign in the second law of thermodynamics (*2*), regulates the direction of what are termed natural or irreversible changes[3]. As will appear later, however, it yields little information about the rate of relaxation towards equilibrium; nor can it predict the particular route of relaxation out of several possibilities which may exist. Perhaps the simplest example of an irreversible process concerns the isothermal expansion of a perfect gas thermally insulated from its surroundings. We suppose the volume of the container to be suddenly expanded from V to $V+\mathrm{d}V$, the external pressure being reduced in accordance with Boyle's law so as to maintain T constant. The gas molecules will at once irreversibly occupy the extra available volume, until they are spread uniformly throughout the whole volume. Now bearing in mind that the same change of state may be achieved reversibly by the method of the last paragraph, we see that dS for the present process is provided by (*4*). On the other hand, since no heat flows in or out of the system, *Equation 3* is simply replaced by the inequality dS is greater than zero, where dS happens to be the net amount of entropy created by virtue of the natural expansion process.

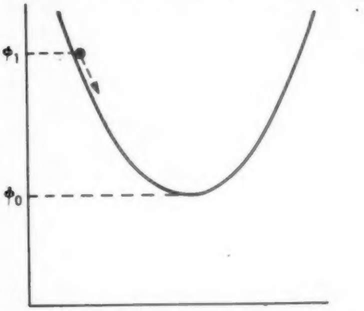


Figure 3. Particle at potential height ϕ_1 will tend to slide down well to equilibrium configuration at ϕ_0

Other examples, such as the melting of ice, can only be discussed qualitatively. Here the entropy function S increases owing to the replacement of long range order, characteristic of a crystal, by short range order characteristic of a liquid. At the same time U increases owing to the rupture and bending of interatomic bonds: if so, neglecting dW, we have

$$\mathrm{d}U = \mathrm{d}Q > 0,$$

implying an entropy flow dQ/T into the system. On balance, $(\mathrm{d}S - \mathrm{d}Q/T)$ exceeds zero above 0°C.

It is important to understand that S may decrease during an irreversible process, provided sufficient entropy flows into the surroundings to provide a compensating increase. Consider the self-recovery of a cold-worked metal. Here S decreases owing to the restoration of lattice perfection; at the same time, the relief of strain energy implies that

$$\mathrm{d}U = \mathrm{d}Q < 0, \qquad ..(5)$$

showing that entropy flows into the surroundings. We thus have dS negative and $(-\mathrm{d}Q/T)$ positive, with $(\mathrm{d}S - \mathrm{d}Q/T)$ positive on balance.

As a second example, crystallizations from the vapour must always involve an enormous decrease in entropy. Part of the decrease arises from the volume contraction on solidification and this is balanced by an entropy flow into the surroundings by much the same mechanism as in the reversible isothermal compression of a perfect gas. The second part of the decrease arises from the build up of configurational order and this is compensated by the formation of interatomic bonds, resulting in a relation of the form (*5*). Accordingly, whilst $\mathrm{d}S$ is large and negative, $(-\mathrm{d}Q/T)$ is large and positive, with $(\mathrm{d}S-\mathrm{d}Q/T)$ positive on balance. It has often been remarked that the growth of biological organization implies a decrease in entropy: all the available evidence indicates that an appropriate gain in entropy occurs elsewhere[4,5]. If so, it must be accepted that the existence of living organisms does not contravene thermodynamic principles.

Mathematical Formulation

To make further progress, we write

$$\mathrm{d}S - \mathrm{d}Q/T = -\mathrm{d}\phi/T \geqslant 0, \qquad \ldots.(6)$$

where ϕ stands for a thermodynamic function having the dimensions of energy. The introduction of this function brings two advantages: it replaces increase in entropy by the more familiar concept of decrease in energy, and the quantity $|\,\mathrm{d}\phi\,|$ in some sense measures the driving energy for the reaction at any stage, since $\mathrm{d}\phi$ equals zero at equilibrium. A valuable analogy for non-equilibrium systems is provided by a particle placed at the side of a potential well (*Figure 3*). Just as the particle tends to slide down the well, thereby diminishing its potential energy until equilibrium is achieved, so a non-equilibrium system tends to pass through a sequence of non-equilibrium states of gradually diminishing ϕ. If a system starts at ϕ_1 and ends up in equilibrium at ϕ_0, the total driving energy for the transition is $\phi_1 - \phi_0$.

An explicit expression for ϕ depends on the circumstances. In the case of a single homogeneous phase, its thermodynamic behaviour is completely controlled by a pair of independent variables T,P or T,V. With the former pair, ϕ becomes identified as the Gibbs free energy G, or thermodynamic potential,

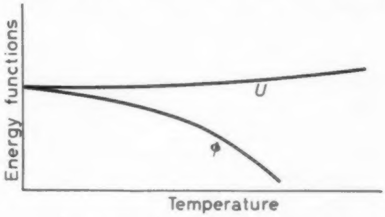


Figure 4. Typical dependence of free energy ϕ and internal energy U on temperature (schematic)

$$G = U - TS + PV \qquad \ldots.(7)$$

of the system. Here U, S, V are all, in principle, functions of temperature and pressure, and therefore vary if the latter vary. Accordingly, for given increments $\mathrm{d}T$ and $\mathrm{d}P$, straightforward differentiations and collecting of terms gives

$$\mathrm{d}G = \mathrm{d}U + P\mathrm{d}V - T\mathrm{d}S - S\mathrm{d}T + V\mathrm{d}P \quad \ldots.(8)$$
$$= (\mathrm{d}Q - T\mathrm{d}S) - S\mathrm{d}T + V\mathrm{d}P, \quad \ldots.(9)$$

on bearing in mind that $\mathrm{d}U+P\mathrm{d}V$ equals $\mathrm{d}Q$.

Coupling *Equations 2* and *9*, we draw the following conclusions. Firstly that for equilibrium at constant temperature and pressure,

$$\mathrm{d}G = \mathrm{d}Q - T\mathrm{d}S = 0$$

and secondly for irreversible processes at constant temperature and pressure

$$\mathrm{d}G = \mathrm{d}Q - T\mathrm{d}S < 0.$$

Strictly speaking, this inequality refers to a non-equilibrium function G which has not yet been explicitly defined. It is supposed, however, that the expression (*7*) still holds, but that U, S and V depend not only on temperature and pressure but on additional parameters relating to the processes taking place. For reversible processes involving varying temperature and pressure, the equation for $\mathrm{d}G$ becomes

$$\mathrm{d}G = -S\mathrm{d}T + V\mathrm{d}P.$$

As a corollary, the equilibrium V corresponding to a prescribed temperature T and pressure P is determined by the 'equation of state'

$$V = \left(\frac{\delta G}{\delta P}\right)_T. \qquad \ldots.(10)$$

For irreversible processes involving a varying temperature and pressure,

$$\mathrm{d}G < -S\mathrm{d}T + V\mathrm{d}P.$$

If the independent variables are T and V we identify ϕ as the Helmholz free energy, or free energy

$$F = U - TS \qquad \ldots.(11)$$

of the system. For given increments $\mathrm{d}T$ and $\mathrm{d}V$,

$$\mathrm{d}F = \mathrm{d}U - T\mathrm{d}S - S\mathrm{d}T$$
$$= \mathrm{d}U + P\mathrm{d}V - T\mathrm{d}S - S\mathrm{d}T - P\mathrm{d}V$$
$$= (\mathrm{d}Q - T\mathrm{d}S) - S\mathrm{d}T - P\mathrm{d}V. \quad \ldots.(12)$$

From *Equations 2* and *12*, it may be concluded that for equilibrium at constant temperature and volume

$$\mathrm{d}F = \mathrm{d}Q - T\mathrm{d}S = 0$$

and, in the case of irreversible processes at constant temperature and volume,

$$\mathrm{d}F = \mathrm{d}Q - T\mathrm{d}S < 0.$$

As before, this inequality refers to a non-equilibrium F formally defined by (*11*), but where U and S depend not only on temperature and volume but on additional parameters relating to the processes taking place. For reversible processes involving varying temperature and volume,

$$\mathrm{d}F = -S\mathrm{d}T - P\mathrm{d}V.$$

As a corollary, the equilibrium P corresponding to a prescribed temperature and volume is determined by the equation of state

$$P = -\left(\frac{\delta F}{\delta V}\right)_T \qquad \ldots .(13)$$

and this equation must, of course, be equivalent to *Equation 10*. Finally for irreversible processes at varying temperature and volume

$$\mathrm{d}F < -S\mathrm{d}T - P\mathrm{d}V.$$

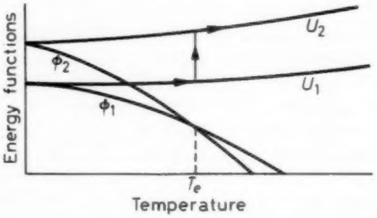


Figure 5. Free energy and internal energy temperature dependence of face-centred cubic lithium (ϕ_1, U_1) and body-centred cubic lithium (ϕ_2, U_2). The two free energy curves cross at a transition temperature T_e in the neighbourhood of 80°K. Note the jump from U_1 to U_2 at T_e (schematic)

Phase Transformations

For most changes involving only solids, the product PV remains practically constant in comparison with U and TS, so that no effective distinction arises between F and G. Alternatively expressed, dW is small by comparison with dQ, so that dU approximates to dQ; if so, the relation (*6*) becomes

$$\mathrm{d}S - \mathrm{d}U/T = -\mathrm{d}\phi/T \geqslant 0,$$

whence $$\phi = U - TS \qquad \ldots .(14)$$

for either temperature and pressure or temperature and volume as the independent variables. Now bearing in mind that neither U nor S depend very critically on temperature, we draw the following qualitative conclusions from *Equation 14:*

(*i*) ϕ approximates to U as T approaches zero, since TS then becomes negligible compared with U and ϕ approximates to $(-TS)$ as T gets large, since TS then dominates U. As a corollary, the dependence of ϕ on T must be as shown in *Figure 4*.

(*ii*) Mathematically, ϕ is minimized at a given temperature and volume, or temperature and pressure, by making U as small as possible and S as large as possible. From the physical point of view, however, these are seen to be mutually incompatible requirements: a small U implies mechanical stability, whereas a large S implies configurational disorder. In practice, the difficulty is overcome by the appearance of structures which emphasize the dominant factor at the temperature concerned.

(*iii*) Taking (*i*) and (*ii*) together, we see that mechanical stability is preferred at low temperatures, hence the appearance of solids. At high temperatures, configurational disorder is preferred, hence the existence of gases. At intermediate temperatures, the two factors are more evenly balanced, resulting in the existence of the liquid phase as a genuine compromise between them.

If a substance has a choice of two different crystalline modifications, of slightly differing free energies (allotropy, polymorphism), the more close-packed structure will be preferred at low temperatures and the more open structure as the temperature increases. These qualitative considerations led C. Zener to postulate the existence of face-centred cubic lithium and sodium at sufficiently low temperatures, a fact subsequently discovered by C. S. Barrett and coworkers. Denoting the face-centred cubic and body-centred cubic free energy curves by ϕ_1 and ϕ_2 respectively, their relative behaviour is indicated in *Figure 5*: both have the same general form, but cross at a transition temperature T_e at which ϕ_1

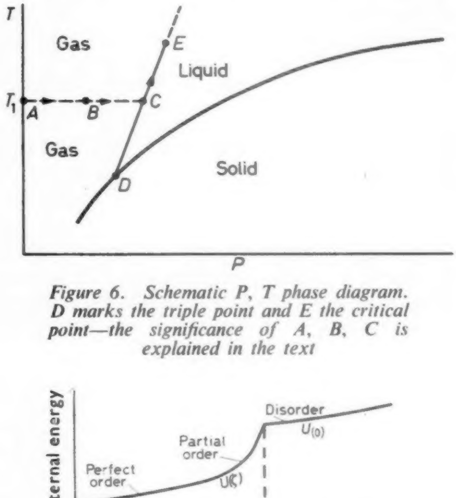


Figure 6. Schematic P, T phase diagram. D marks the triple point and E the critical point—the significance of A, B, C is explained in the text

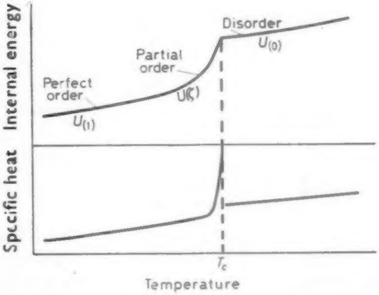


Figure 7. Temperature dependence of internal energy and specific heat in neighbourhood of critical temperature T_c for order–disorder transition (schematic)

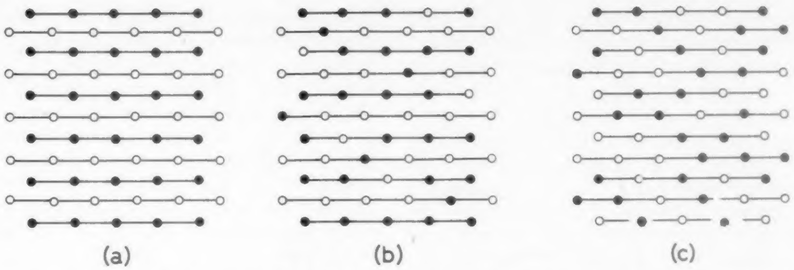


Figure 8. (a) Model of perfect order: successive lattice rows are alternatively black and white; (b) partial order: successive lattice rows are alternatively mainly black and mainly white; (c) Disorder: random distribution of black and white in each row

equals ϕ_2, *i.e.*,

$$U_1 - T_e S_1 = U_2 - T_e S_2, \qquad \dots.(15)$$

and $$L = U_2 - U_1 = T_e(S_2 - S_1) > 0, \qquad \dots.(16)$$

where L stands for the latent heat of transformation. The relation *(16)* governs the way in which U and S jump at the transition temperature, it being noted that both become greater for the high temperature phase.

Quantitative calculations of ϕ_1 and ϕ_2 for sodium, undertaken very recently by Dr M. A. E. NUTKINS[6] on the basis of Debye's theory of solids, have provided a theoretical check on Zener's ideas and a transition temperature of the right order of magnitude. At temperatures in the neighbourhood of the transition temperature T_e, we have the useful approximation

$$\phi_2 - \phi_1 = (U_2 - U_1) - T(S_2 - S_1) = L - TL/T_e$$
$$= L(1 - T/T_e), \qquad \dots.(17)$$

for the free energy difference between the two phases. Similar considerations apply to other polymorphic transitions, and to the solid–liquid transition. Solid–gas and liquid–gas transformations involve such enormous volume changes that the Gibbs free energy G differs significantly from the Helmholz free energy F, thus making it necessary to specify whether temperature and volume or temperature and pressure are the controlling variables; otherwise, the preceding general ideas still hold good.

It is usual to assemble information about phase relations in a pressure–temperature diagram as typified in *Figure 6*. The field of representative points divides into a number of regions, each characterized by a specific equilibrium phase; two phases co-exist in equilibrium along the boundary between them, and three phases co-exist at the triple point. The liquid and gas phases are conventionally distinct only up to the critical point, at which their specific volumes become equalized. Recent work, however, suggests that the boundary should be continued as indicated by the dashed line, the transition across the latter being of a type discussed in the next section[7].

As an exercise in the use of the phase diagram, let us examine what happens to a quantity of liquid placed in an evacuated rigid container held at a temperature T_1. Not being subjected to any pressure initially, the representative point A lies initially on the T-axis at a height T_1: from the diagram we see that gas is the equilibrium phase at A, and hence deduce that the liquid irreversibly evaporates. As evaporation proceeds, P necessarily increases, and the representative point therefore moves parallel to the P-axis along the dotted line as shown. There are now two possibilities. The liquid has been completely vaporized at a point B within the gas phase field. If so, the system at B has the status of a gas of fixed volume (that of the container) currently at a temperature T_1. Alternatively, the liquid has not been completely vaporized before the phase boundary is reached at C. If so, liquid and vapour co-exist in equilibrium at C, and transformation has no tendency to proceed further. On raising the temperature beyond T_1, the representative point moves along the phase boundary in the direction of increasing pressure, indicating further evaporation.

Second Order Transitions

The transitions so far discussed are termed first-order since they involve a discontinuity in the internal energy U. There exists a class of transformations which involve no discontinuity in U, only in the specific heat $\mathrm{d}U/\mathrm{d}T$, and which are hence said to be second order. Sometimes they are termed lambda-transitions, in view of the characteristic shape of the specific heat curve at the critical temperature T_c (*Figure 7*). Perhaps the simplest example of such a transition is the order–disorder change in binary alloys[8]. Below T_c, the two kinds of atom are arranged perfectly regularly on a common lattice; as T approaches T_c, the regularity begins to disappear, at first gradually and then catastrophically, until just beyond T_c the distribution has become completely random (*Figure 8*).

To describe this situation within the framework of classical thermodynamics, we introduce an order parameter ζ, having the same status as P, V or T: perfect order is defined by ζ being unity, perfect disorder by ζ being zero, and states of partial order by intermediate values of ζ. As the temperature varies in the critical range, ζ varies continuously from one of its extreme values to the other, as sketched in *Figure 9*. It is convenient here to think of each intermediate structure, for example that characterized by ζ_1, as existing hypothetically over the whole temperature range, with a free energy function

$$\phi(T,\zeta_1) = U(\zeta_1) - TS(\zeta_1) \qquad ..(18)$$

that depends on temperature in the usual way. We thus set up a one-parameter family of curves arranged as in *Figure 10*, the members of which respectively at successive temperatures make up the equilibrium free energy curve of the system. From this point of view the terminal structures zero and one have essentially the same status as ζ_1, only the latter appears as the equilibrium phase at just one specific temperature T_1.

To calculate T_1, we argue as follows. The phase ζ_1 formally undergoes an infinitesimal first order transition to $\zeta_1 + \mathrm{d}\zeta_1$ in the neighbourhood of T_1, whence

$$\phi(T_1,\zeta_1 + \mathrm{d}\zeta_1) = \phi(T_1,\zeta_1),$$

i.e.,

$$\frac{\mathrm{d}\phi}{\mathrm{d}\zeta_1} = \frac{\mathrm{d}U}{\mathrm{d}\zeta_1} - T_1\frac{\mathrm{d}S}{\mathrm{d}\zeta_1} = 0. \qquad ....(19)$$

Since $U(\zeta_1)$ and $S(\zeta_1)$ are known, *Equation 19* yields

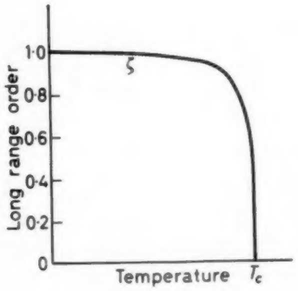


Figure 9. Temperature dependence of long range order parameter

T_1 as a function of ζ_1, and hence also the equilibrium free energy of the system $\phi(T_1,\zeta_1)$ at T_1. In the usual formulation of the theory, emphasis is placed on a particular temperature Θ, and the minimum free energy at Θ selected from the set

$$\phi(\Theta,1), \ldots \phi(\Theta,\zeta_1), \ldots \phi(\Theta,0)$$

by the equation

$$\phi'(\Theta,\zeta) = \frac{\mathrm{d}U}{\mathrm{d}\zeta} - \Theta\frac{\mathrm{d}S}{\mathrm{d}\zeta} = 0,$$

which provides the equilibrium ζ in terms of Θ. Evidently this formulation is quite equivalent to the preceding, the former being perhaps philosophically superior in enabling the second order transition to be identified as a continuous series of infinitesimal first order transitions. It may be remarked that the latent heat of transformation at any stage is $U'(\zeta)\mathrm{d}\zeta$, so that the total latent heat is

$$\int_{\zeta=1}^{\zeta=0}\frac{\mathrm{d}U}{\mathrm{d}\zeta}\,\mathrm{d}\zeta = \int_{T_c=T_c-\Delta T}^{T=T_c}\frac{\mathrm{d}U}{\mathrm{d}T}\mathrm{d}T.$$

This is absorbed over a finite temperature interval, and hence identified as a catastrophic increase in the specific heat.

Notwithstanding that the product PV may be neglected in comparison with U and TS, first order transformations are always accompanied by a macroscopicallly discernible change in volume. This feature, indeed, sometimes provides the most direct indication that a transformation has taken place. By contrast, second order transformations do not involve a volume change since the atoms merely undergo redistribution over a common lattice. The use of the term second order is seen to be entirely consistent with the greater subtlety inherent in the latter situation.

The Activation Energy Barrier

Most irreversible processes inevitably bring into play forces which, to a greater or lesser extent, oppose the reaction. At low temperatures, and sometimes even at room temperatures, these may completely inhibit the approach to equilibrium, as with glass or quenched-in lattice defects. At higher temperatures, however, the thermal energy of the atoms suffices for them to surmount the ' activation energy barrier' and thus enable the reaction to proceed. A useful picture of the situation is provided by the single particle model. Instead of the side of the well being perfectly smooth, as envisaged in *Figure 11*, we may picture it more realistically as broken by local hollows and humps which impede descent.

Suppose the particle rests in local equilibrium at a potential level ϕ, and that this is separated from the next level of local equilibrium at $\phi+\mathrm{d}\phi$ by a potential hump of height A. Although potential energy can be reduced by the transition from ϕ to $\phi+\mathrm{d}\phi$, no direct mechanism exists for overcoming the barrier A separating the two configurations. However, an indirect mechanism is provided by thermal vibrations. In effect, these allow the particle to occupy a series of harmonic oscillator levels 0, $\varepsilon_1, \varepsilon_2, \dots$, with a relative chance $e^{-A/kT}$ of acquiring sufficient thermal energy to jump the barrier. Sooner or later this latter contingency will be realized, and the particle descends to the next configuration $\phi+d\phi$, so continuing until final equilibrium has been achieved.

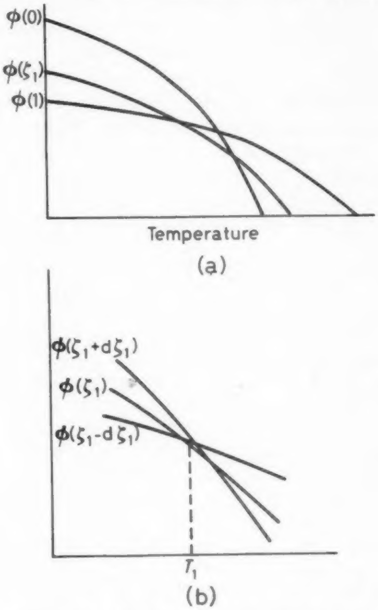


Figure 10. (*a*) *Free energy curves* $\phi(1)$, $\phi(\zeta_1)$, $\phi0)$, *referring to structure of perfect order, partial order* ζ_1 *and complete disorder, respectively;* (*b*) *Close-up view of situation in neighbourhood of temperature* T_1, *where* $\phi(\zeta_1)$ *crosses* $\phi(\zeta_1-d\zeta_1)$ *and is in turn crossed by* $\phi(\zeta_1+d\zeta_1)$

It may be noted that $e^{-\varepsilon/kT}$ approaches unity when the temperature T gets large, in other words each level ε tends to have an equal chance of being occupied, with the result that the higher energy levels are favoured at the expense of the lower. Consequently, the particle acquires a progressively increasing chance of surmounting the barrier. This conclusion forms the quantitative basis for understanding the dominant influence of temperature on the rate of reaction.

The preceding ideas have an immediate application to the theory of diffusion, particularly the diffusion of solute atoms through a crystal. Any solute atom lies within a potential well of height q, defined by the periodic field of the crystal (*Figure 12*), with a relative chance $e^{-q/kT}$ of jumping into a neighbouring position. On multiplying this chance by ν, the frequency of atomic vibrations, we arrive at $\frac{1}{3}\nu e^{-q/kT}$ for the number of successful jumps per second in a particular direction, bearing in mind that there are three equally likely independent directions of vibration. Typically, q is of the order one electron-volt, kT of order one fortieth of an electron-volt at room temperature, and ν of order 10^{13} sec^{-1}. If now there are n solute atoms per unit volume, concentrated in lattice planes of spacing 'a' in the diffusion direction, this implies na solute atoms per unit area of lattice plane: $\frac{1}{2}na$ of these will, on the average, attempt to jump forward, of which

$$\tfrac{1}{2}na.\ \tfrac{1}{3}\nu e^{-q/kT} \qquad \dots.(20)$$

per unit time will be successful. The diffusion coefficient is usually defined to be

$$D = \tfrac{1}{6}\nu a^2\, e^{-q/kT}, \qquad \dots.(21)$$

in terms of which (*20*) may be written nD/a. By measuring the variation of the diffusion coefficient D with temperature, or rather log D against $1/T$, we may form an estimate of q and ν. It will be noted that the diffusion coefficient depends on the particular direction under consideration.

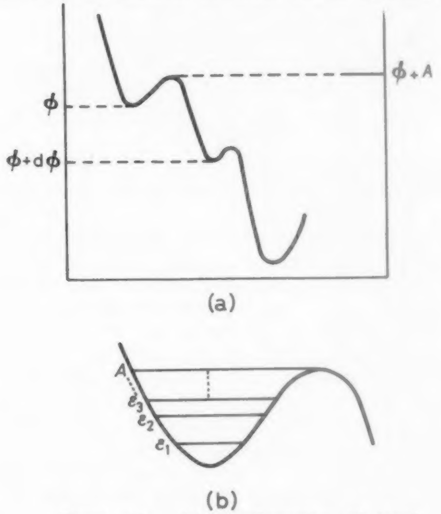


Figure 11. (*a*) *Configuration of local equilibrium* ϕ *separated from neighbouring configuration* $\phi+d\phi$ *by a local barrier A;* (*b*) *Close-up view of situation at* ϕ. *By virtue of the thermal levels* 0, ϵ_1, ϵ_2 . . . *A* . . . *the particle has a certain chance of surmounting the barrier*

We now extend the argument to allow for the effect of an external potential field ϕ interacting with the solutes, *e.g.*, as in A. H. Cottrell's theory of atmosphere formation around dislocations. Supposing ϕ to decrease slowly over a few interatomic spacings in the diffusion direction, its effect will be to damp the amplitude of the periodic crystal field in that direction, and so impose a statistical directionality on the otherwise random thermal agitation of the atoms. Quantitatively formulated, if the activation energy barrier diminishes from $q+\phi_1$ to $q+\phi_2$ over an interatomic spacing as shown (*Figure 12*), the excess relative chance of jumping down the gradient is

$$e^{-(q+\phi_2)/kT}-e^{-(q+\phi_1)/kT}$$

which equals

$$\frac{\phi_1-\phi_2}{kT}.e^{-q/kT}=\frac{a}{kT}.\frac{d\phi}{dx}.e^{-q/kT} \qquad ..(22)$$

on bearing in mind that $\phi_1-\phi_2$ is equal or nearly equal to $-a.d\phi dx$ and assuming ϕ to be small in comparison with kT.

Multiplying (*22*) by the same factors as previously, we arrive at

$$-\frac{nD}{kT}.\frac{d\phi}{dx}$$

for the net number of solutes per unit area drifting down the potential gradient per unit time. The 'drift velocity' is readily seen to be given by

$$v=-\frac{D}{kT}.\frac{d\phi}{dx} \qquad ..(23)$$

a formula originally deduced by Einstein in his theory of Brownian motion, though with a mobility factor arising from molecular collisions instead of from a periodic lattice potential. It should be noted that D/kT approaches zero both when T is either large or small: at high temperatures the solutes readily jump the barrier in either direction, regardless of the perturbing potential ϕ, so that v has zero value on the average; at low temperatures the solute atom remains trapped in its potential well, so that again v is zero. At intermediate temperatures the combination of the two opposing factors, namely the energy barrier and thermal fluctuations, produces the effect of a viscous drag impeding the motion of the solute and imparting to it an average limiting velocity given by *Equation 23*.

Similar arguments show that $-D.dc/dx$ gives the number of solutes per unit area drifting per unit time down a concentration gradient. The corresponding drift velocity is

$$\frac{-D}{c}.\frac{dc}{dx}.$$

Metastability

Cases often arise where the activation energy barrier is so high as to be completely immune to thermal fluctuations. The system then carries on existing in the relevant state of local equilibrium, with no prospect of ever achieving the greater stability potentially available. This phenomenon is termed metastability (metastable equilibrium). The situation in terms of the one particle model appears as indicated in *Figure 13*. Perhaps the most outstanding example of a metastable system is diamond, which co-exists at all temperatures with the more stable modification graphite (*Figure 14*).

The thermodynamic status of diamond differs in several important respects from that of a non-equilibrium system such as glass. Some of the differences can be appreciated directly by comparing the respective one particle models: diamond has a clearly defined internal structure, whereas glass has not; diamond remains locally stable against any thermal fluctuations that may occur, whereas glass is locally unstable against thermal fluctuations that lead to increasing order. Diamond, in contrast to glass, obeys an equation of state and has a vanishing entropy at 0°K. If graphite did not exist, diamond would be regarded as the stable equilibrium state of crystalline carbon over the whole temperature range of its existence. This example makes it clear that stable and metastable equilibrium are purely relative terms, their use depending on the energy level of the reactions under consideration: no genuine distinction can be drawn between the two, and all the principles of thermodynamics, *e.g.*, the third law, have the same validity for the one as for the other.

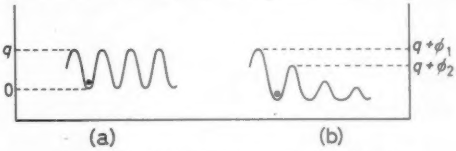


Figure 12. (a) Periodic crystal field; (b) effect of slowly varying perturbing potential ϕ

Metastable systems abound in nature, indeed they comprise most of the systems of actual experience. Diamond has already been mentioned. Martensite, the main hardening constituent of steels, is metastable relative to a mixture of ferrite and cementite. Most metals are metastable, *i.e.*, the mixture metal plus oxygen is metastable relative to the metallic oxide. Supercooled liquids are metastable relative to the crystalline phase, and supersaturated vapours are metastable relative to the liquid or crystal. A mixture of hydrogen and oxygen is metastable relative to water. On the microscopic scale, metals owe their mechanical properties to metastable lattice defects known as dislocations. Organic molecules are weakly metastable, thus enabling energy to be trapped and made available for use by the organism when required. As regards the kinetics of chemical change, Sir C. N. Hinshelwood[9] has remarked that 'the real unities underlying the interpretation of chemical changes are seldom open and apparent but tangled in a confusion which here and there might well appear inextricable, but did not a broader view reveal that through the tangle run certain threads bright enough to show a plan. The strongest and brightest of these threads is the idea of the activation energy'.

Whence arises the activation energy barrier which maintains the various metastable systems enumerated? An instructive answer is provided by considering the mechanism of phase nucleation in condensed systems, *e.g.*, that of a crystal from supercooled liquid. According to the generally accepted picture, we envisage embryos of crystal phase created by random statistical fluctuations of atomic structure within the liquid. These embryos are subject to two energy factors, the first is the driving energy for the transformation from liquid to crystal, *i.e.*, the free energy difference, ϕ, per unit volume between crystal and liquid which is markedly temperature dependent: at the equilibrium temperature T_e between liquid and crystal, ϕ has the value zero; above T_e it is positive, and below T_e it is negative. The second is the surface energy σ per unit area created by the disregistery of structure at the boundary of the embryo: this is always positive and largely independent of T. Accordingly, the energy required to create an embryo of radius r at temperature T is given by

$$a = a\,(r,T) = \phi r^3 + \sigma r^2$$

apart from unimportant numerical factors. At any temperature T above T_e, this function increases steadily with r so that the embryo is always unstable. For any T below T_e, the function at first increases but passes through a maximum

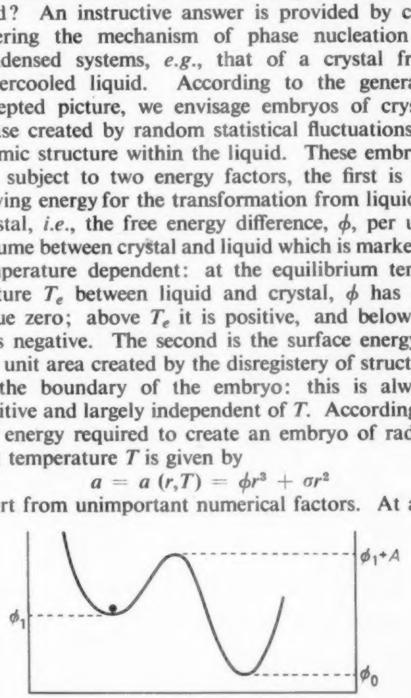


Figure 13. Metastable equilibrium. The particle rests in local equilibrium at ϕ_1, with no chance of surmounting the barrier A separating it from the configuration ϕ_0 of greater stability

$$A = A(T) = \sigma^3/\phi^2 \qquad ..(24)$$

and at a critical radius

$$R = R(T) = -\sigma/\phi, \qquad ..(25)$$

as sketched in *Figure 15*. This maximum defines the activation energy which must be supplied to the embryo before it can qualify as a stable nucleus of crystal phase. It is the presence of the barrier A which ensures that supercooled liquid is everywhere in local internal equilibrium.

Bearing in mind that $|\phi|$ increases with the degree of supercooling, we can see from (*24*) and (*25*) that A and R both decrease as the temperature decreases. The equilibrium number of nuclei at temperature T is provided by the Boltzmann factor $e^{-A/kT}$; arguments can be advanced to show that A/T decreases, and hence the Boltzmann factor increases as T diminishes. Accordingly, as the temperature is lowered below T_e, two possibilities will arise. At least one stable nucleus is created almost at once by virtue of a fluctuation and having reached the critical size then grows irreversibly until the phase change has been completed. Alternatively, no stable nucleus is created until a temperature T' well below T_e: within this temperature interval, the liquid is said to be supercooled, having then the status of a metastable system relative to the crystalline phase.

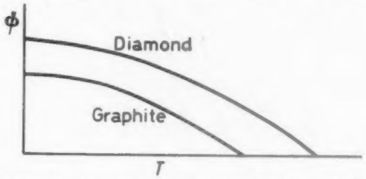


Figure 14. Free energy curves of graphite and diamond (*schematic*)

The above analysis envisages nucleation as a random process, having an equal chance of taking place anywhere throughout the medium. In practice, however, it is usually triggered off at impurities, at defects or at boundaries: these provide concentrations of available energy, which readily enable the requisite local fluctuations in structure to be effected. Indeed, without such preferred sites, supercooling and supersaturation would become much more pronounced phenomena than is actually the case. For instance water usually freezes at 0°C, but by careful control of size and impurities[10] it can be supercooled through as much as 40°C. According to Frank's theory of crystal growth from the vapour, crystallization only takes place at the observed rate because suitable nucleation centres are provided by dislocations. Gas reactions are usually initiated at cracks and cavities in the walls of the container. The two mechanisms of nucleation are termed homogeneous and heterogeneous respectively.

For reactions within solids, the behaviour of a nucleus is controlled by a third factor, the elastic strain energy, besides the other two. This arises because the product phase usually has a different specific volume from that of the matrix, thus generating accommodation stresses in the latter. Strain energy, unlike the surface energy factor, is proportional to the volume, and could therefore remain important even when the nucleus has grown into a precipitate of macroscopic dimensions. This applies particularly to martensite plates, owing to the coherency requirements at the boundary between matrix and product. The thermodynamics of martensite systems provides a fascinating example of interplay between chemical and mechanical energy, resulting in such phenomena as rubber-like elasticity and thermo-elastic equilibrium. More is likely to be heard about these aspects in the future, but they cannot be pursued further within the limits of the present article.

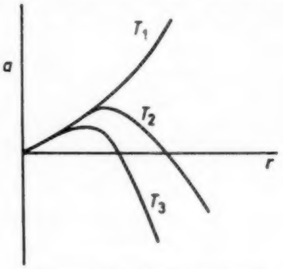


Figure 15. $T_1 > T_e > T_2 > T_3$

The Thermodynamic Field

We now introduce the concept of the thermodynamic field. This may be pictured as a medium where the thermodynamic variables, such as pressure and temperature, depend on position in a prescribed way. Accordingly,

$$P = P(x, y, z) \text{ and } T = T(x, y, z)$$

where x, y, z denote the coordinates of a point. A small system of given phase structure thus has a free energy

$$\begin{aligned}\phi = \phi(P,T) &= \phi[P(x, y, z), T(x, y, z)]\\ &= \phi(x, y, z), \qquad ..(26)\end{aligned}$$

which varies with position, and will therefore tend to drift through the medium towards the nearest local minimum of ϕ. By analogy with the drift of solute atoms through a crystal, we write

$$v = -\frac{D}{kT}\nabla\phi.$$

For most macroscopic systems, either D or $\nabla\phi$ is so small that the system remains fixed at some point, in stable or metastable equilibrium relative to the variables P and T operative at the point, but formally in non-equilibrium relative to the thermodynamic field as a whole.

An interesting example of macroscopic finite drift occurs with servo-mechanism devices. Here the system contains a store of potential energy coupled to the local field conditions, and released according to some specified law, for example, so as to propel the system in the direction of decreasing temperature. Identifying ϕ of (*26*) as the stored potential energy, the servo-mechanism may be regarded as a system moving in a resisting medium under the influence of a field of force ϕ. Much of the theory of servo-mechanisms (cybernetics)[11] is taken up with the way information is signalled from environment to system, and with how the system responds. An ordinary dynamical system differs from a servo-mechanism only in the fact that it is coupled directly to the field, *e.g.*, gravity, in a formal way that cannot be understood on the basis of any physical picture. Elementary biological organisms may be identified as servo-mechanisms in the first approximation, the energy store being chemical rather than electro-mechanical. Their behaviour has been variously described as goal-seeking, purposeful teleological, homeostatic. However, within the framework of thermodynamic principles, nothing more is involved than drift down a free energy gradient.

References

[1] BERNAL, J. D., *Nature, Lond.* 183 (1959) 141
[2] DAVIES, R. O. and JONES, G. O. *Phil. Mag. Supp.* 2 (1953) 370
[3] — *Science News 28*, p. 41 London: Penguin, 1953
[4] BUTLER, J. A. V. *Science News 18*, p. 21 London: Penguin, 1950
[5] KLEIN, M. J. *Brit. J. Phil. Science* 4 (1953) 370
[6] NUTKINS, M. A. E. *Proc. Phys. Soc.* 72 (1958) 810
[7] JONES, G. O. *ibid.* 69 *B* (1956) 1348
[8] DOMB, C. *Science Progress* 43 (1955) 402
[9] HINSHELWOOD, C. N. *Kinetics of Chemical Change*, p. 268 Oxford: Clarendon, 1940
[10] MASON, B. J. *The New Scientist* (1957)
[11] CHERRY, E. C. *Nature, Lond.* 172 (1953) 648

GENERAL

COTTRELL, A. H. *Theoretical Structural Metallurgy* London: Arnold, 1955

TEMPERLEY, H. N.V. *Changes of State* London: Cleaver-Hume, 1956

SMOLUCHOWSKI, R. *et al. Phase Transformations in Solids* New York: Wiley, 1951

SLUCKIN, W. *Minds and Machines* London: Penguin, 1954

SURVEY

Recording Rail Track Behaviour Under Dynamic Loading

It is axiomatic that before passenger trains can be driven at high speed the track on which they are to run must be laid accurately and maintained in that condition. Early track recording gear which was of limited value has now been superseded, so far as British Railways are concerned, by a new coach recently completed by Elliott Brothers (London) Ltd. The coach has been under development for about two years, though a design study was commenced earlier to facilitate preparation of a specification.

The new coach provides means for assessing track conditions under a loaded vehicle while it is moving at speed along the track; this enables track to be surveyed while normal timetables are in operation, and British Railways stipulated that measurements should be made in such a way as to make them independent of the movement of the vehicle. The coach has been designed as a four-wheeled self-propelled vehicle, operable from either end, to record at any speed up to 30 miles per hour and when not recording to travel at speeds up to 55 miles per hour. One axle only is driven; the other is used to provide mechanical drive to the distance-measuring apparatus.

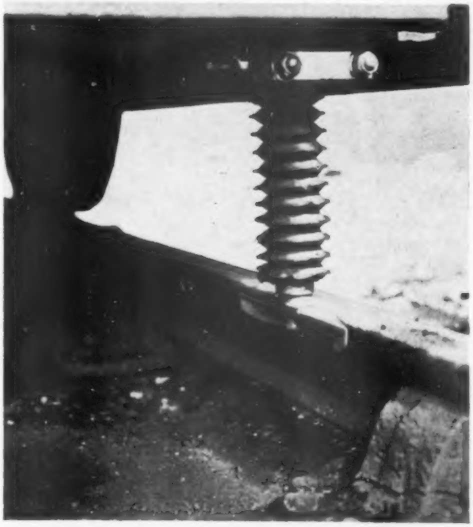


Figure 1. The Elliott track recording coach —the photograph illustrates the sensing shoe along the running rail

Continuous records (traces on photosensitive paper) of superelevation (cant), curvature and gauge are produced, and additional equipment is provided so that a record of an earlier journey by the coach over the same track can be run through in synchronism. This enables engineers to see immediately whether track adjustments which may have been made in the interim have had the desired effect. Compensating adjustments can be made on the recording apparatus to take account of the continuous rotation of the earth (every half-hour) or the motion of the coach around curves. This is necessary because the datum for cant measurement is provided by a high speed gyroscope independently of curvature or gradient: this is compared with the tilting of one axle. Curvature and gauge are obtained by measuring the movements of a system of probes which contact the inside edge of the head of the rail.

These probes (*Figure 1*) are designed to negotiate track fittings at the maximum recording speed. The rubbing faces of the sensing shoes are of Stellite, and it is claimed that wear is so small that no allowance need be made for this even on tests exceeding 1000 miles in length. It should be noted that the cone to which the running wheels of this coach are turned is 1:100. Normal wheels would introduce too much variability into the record of cant. Normal loading is about 12 ton/axle, and the coach carries its own generating equipment to supply the gyroscopes and synchros at 400 c/s, three phase. Living accommodation for the crew is also provided.

Industrial Application of Fluid Flow

The British Hydromechanics Research Association recently held Open Days at their establishment at Harlow, Essex. This Association is non-profit making, being controlled and financed by members with the help of a proportionate government grant. Any industrial applications of fluid flow within fixed boundaries is of interest to the Association and problems investigated include: mine ventilation; centrifugal pump design; water hammer in pipelines; seals, packings and joints; cavitation; hydraulic transport of solids; friction losses in pipe fittings; and the hydraulic design of civil engineering structures. In fact, during work on the last-named problem, it has proved possible so to modify channel design and control valve equipment as to make large savings of capital on initial outlay and to reduce running costs substantially by eliminating turbulence, removing potential vortices and reducing frictional and head losses in various ways.

BOOK REVIEWS

A History of Western Technology

F. KLEMM

(401 pp; 8¾ in. by 5½ in.)

London: George Allen & Unwin. 32*s*

THERE are various ways in which one might attempt to write a history of technology. For instance, one could describe the principles involved in the development of machines and techniques; one might discuss the individual inventors, and innovators of techniques, giving their biographies and an assessment of the importance of their work; or one could, as Mumford has done, write a history of society and its dependence on the development of machines.

In writing *A History of Western Technology*, Dr KLEMM has done none of these things. In fact, his book may perhaps be more properly described as an anthology of carefully selected contemporary writings connected by commentary in the text, rather than as a history. Dr Klemm is a man of very wide erudition, and his extracts range from those taken from technical treatises such as *De Re Metallica* (from which he gives a good deal of the most interesting introduction) to quotations from contemporary political and religious writings. His intention appears to be not only to describe the progress of technology but also to outline the background conditions under which this progress occurred. In this aim he succeeds admirably; and at the same time he gives the reader access to material and documents which are not always readily available. It would not, one feels, be going too far to say that this method is ideal if one wishes to outline the progress of technology (or science) and to deal with the subject in a humanistic way.

There is, however, one slight criticism of this otherwise admirable book that one might venture. It is a relatively short book, but in it the author attempts to cover the whole of man's technological development from Grecian times to the detonation of the first atomic bomb. Somewhat naturally, therefore, the quality of the treatment varies. The earlier periods up to and including the Renaissance are very good; but the later periods tend to be treated rather sketchily. One feels that Dr Klemm has tried to cover too much ground and that it would have been better to have ended the book at about 1900 (at the end of what Mumford calls the 'Palaeotechnic Phase') rather than to have ventured into the present phase of extremely rapid technological development.

This is a history of Western, and not British, technology. Inevitably, therefore, there are omissions and inclusions which, from our insular viewpoint, we may find surprising. Some of these are obviously salutory as, for instance, the full discussion of Papin's contribution to the development of the steam engine which is often overshadowed here by the work of Savery and Newcomen. Less excusable in this context, however, is the absence of any reference to Robert Hooke who was no only Papin's guide and mentor for many years but who also made many very important contributions to technology.

Nevertheless, this book can be thoroughly recommended. It is well written, and the translation is extremely good. The plates and diagrams are well chosen and extremely well produced. The list of sources, the bibliography and the index are excellent; and as was said above, the texts quoted have been very carefully chosen and include translations from many which are difficult of access in this country. The book fills a gap which many of us are conscious of on our shelves; it provides information and diagrams which few of us have readily available; but above all it is a notable addition to the all too sparse literature on this most important subject.

H. D. TURNER

Meson Physics

R. W. MARSHAK

(viii+378 pp; 8 in. by 5 in.)

New York: Dover Publications;

London: Constable. 16*s* (paper covered)

THIS book is a republication in cheaper form of a volume first published in 1952 by the McGraw Hill Book Company. In 1952 Professor Marshak's book was welcomed as one of the first critical surveys of the rapidly advancing subject of meson physics. It must now be judged against more recent books, for example, Professor Jackson's *The Physics of Elementary Particles* published in 1958. Comparison makes one doubt the wisdom of republication in an unabridged and unaltered form since progress in this field has been so great. When *Meson Physics* was originally written the contribution of cosmic rays to pion and muon physics was considerable, and to the new particles, K-mesons and hyperons, all important and decisive. But with the coming into operation of the great proton synchrotrons at Brookhaven and Berkeley in 1953 and 1954, and the development of new techniques such as the diffusion chamber and the bubble chamber, the position has radically changed. New particles have been discovered and a wealth of factual information has been obtained.

Advances on the theoretical side since 1952 have been no less impressive. Of especial significance has been the discovery of the non-conservation of parity in weak reactions with its important consequences for the theory of the neutrino and meson decay processes. Mention should also be made of associated production of the 'strange' particles, and the Gell–Mann–Nishijima scheme.

Re-reading Professor Marshak's book after a lapse of six years one is impressed with his success in capturing the drama of the challenging enterprise which is modern particle physics. The drama is still being enacted but those who know of recent progress will wish for Acts I and II and not just Act I.

G. D. ROCHESTER

See also page 273

NOTICE TO CONTRIBUTORS: All communications intended for the Editorial Department should be addressed to The Editor, *Research*, 4 and 5 Bell Yard, London, W.C.2. Contributions sent on approval will be carefully considered by the Editor; but no responsibility whatsoever can be accepted for manuscripts, although every effort will be made to return them if a stamped addressed envelope is enclosed for that purpose. The copyright of all contributions for which a fee is paid shall belong to Butterworths Publications Ltd, together with the right of republication. The Editor reserves the usual right to make alterations in the text of articles accepted for publication. The author will receive galley proofs and if alterations have been made every effort will be made to obtain the author's prior agreement. Unless special arrangements are made, contributions are only received and considered on these conditions.